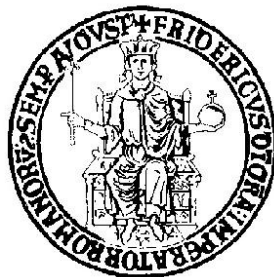


UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II



Facoltà di Ingegneria

Dottorato di ricerca in Ingegneria dei Sistemi Idraulici, di Trasporto e Territoriali
XXIV ciclo: Esame finale nell'Indirizzo "Ingegneria dei sistemi di Trasporto"

Candidato

Egidio Quaglietta

Titolo della Tesi

**A Microscopic Simulation Model For Supporting The Design Of
Railway Systems: Development And Applications**

Coordinatore di dottorato:

Prof. Ing. Guelfo Pulci Doria

Coordinatore di indirizzo:

Prof. Ing. Bruno Montella

Tutor universitario:

Prof. Ing. Vincenzo Punzo

Tutor aziendale:

Ing. Giovanni Bocchetti

Controrelatore:

Prof. Ing. Luigi Biggiero

The research presented in this Thesis has been supported by a Ph.D. grant of the Italian railway systems supplier Ansaldo STS.



Index

Chapter 1. Introduction	1
1.1. Motivations and Objectives of research.....	1
1.2. Thesis description and contributions	2
1.3. Thesis Outline.....	6
Chapter 2. A General Outlook on Railway systems: Infrastructure components and operations.....	8
2.1. Introduction to Time-driven and Event-driven systems	10
2.2. Railway networks as hybrid continuous-discrete systems	13
2.3. Railway infrastructure	16
2.3.1. <i>Terminal Stations.....</i>	<i>19</i>
2.3.2. <i>Ordinary Stations</i>	<i>20</i>
2.3.3. <i>Stop Stations</i>	<i>21</i>
2.4. Timetable	23
2.4.1. <i>Principles of Train separation.....</i>	<i>24</i>
2.4.2. <i>The Blocking Time model</i>	<i>26</i>
2.5. Rolling Stock	31
2.5.1. <i>Basics of rail vehicle motion</i>	<i>33</i>
2.5.2. <i>Active Forces.....</i>	<i>35</i>
2.5.3. <i>Vehicle Motion Resistances.....</i>	<i>40</i>
Traction unit resistances	40
Vehicle resistances for Passenger Trains.....	41
Vehicle resistances for Freight Trains	41
Line Resistances	42
2.5.4. <i>Vehicle Motion Formula and Rotating masses</i>	<i>42</i>
2.6. Signalling equipments and other computer-based systems	43

2.6.1. <i>Automatic Train Protection</i>	44
Group 1: Systems with intermittent transmission and without braking supervision.....	48
Group 2: Systems with intermittent transmission at low data volume and with braking supervision	49
Group 3: Systems with continuous transmission of signal aspects by coded track circuits	49
Group 4: Systems with intermittent transmission at high data volume and dynamic speed supervision.....	50
Group 5: Systems with continuous transmission at high data volume and dynamic speed supervision.....	51
A unified European Train Control Systems: the ETCS levels	52
2.6.2. <i>Automatic Train Operation</i>	57
ATO Stopping, Docking and Starting	58
ATO/ATP systems for Multi-Aspect Signalling in Metro networks.....	60
2.6.3. <i>Interlocking Systems</i>	61
Mechanical interlocking	67
Electro-Mechanical interlocking	68
Relay interlocking	68
Electronic interlocking	68
Chapter 3. An Overview on simulation models of Railway Systems and their applications in practice.	70
3.1. Introduction	70
3.2. Network Modelling.....	70
3.2.1. <i>Macroscopic models</i>	75
The NEMO model	77
The SIMONE model.....	79
3.2.2. <i>Mesosopic models</i>	81

A mesoscopic model for simulating freight train operations	82
3.2.3. <i>Microscopic Models</i>	85
The OpenTrack model	88
The RailSys model	90
3.2.4. <i>Hybrid models</i>	93
3.3. Synchronous simulation models	96
3.4. Asynchronous simulation models.....	97
3.5. Deterministic and Stochastic simulation models	98
3.6. Applications of railway simulation models for supporting design activities.	99
3.6.1. <i>Verifying the stability of timetable and network.....</i>	<i>100</i>
3.6.2. <i>Verifying system performances for an optimized signalling layout</i>	<i>102</i>
3.6.3. <i>Designing of robust timetable</i>	<i>107</i>
3.6.4. <i>Real-time rescheduling of train operations.....</i>	<i>110</i>
Asynchronous simulation	111
Synchronous simulation	112
Optimization methods	112
Chapter 4. Development of a microscopic infrastructure model for simulating railway operations.	115
4.1. Introduction	115
4.2. General features of the simulation model	117
4.3. Infrastructure module.....	120
4.4. Rolling Stock module	124
4.5. Signalling system module.....	129
4.6. Timetable module	137
4.7. Simulation core	139
4.8. Simulation Outputs	145
4.9. Parallelization and its effects on computing efficiency	149

4.10. Model Validation	156
4.11. Uncertainty analysis of the model: Sensitivity Analysis.	159
<i>4.11.1. Sobol’ variance-based method for performing sensitivity analysis.....</i>	<i>162</i>
<i>4.11.2. The case of a MRT system: the Cumana line.</i>	<i>166</i>
Results	170
Chapter 5. Practical applications of the microscopic model to support different design activities.	176
5.1. Introduction	176
5.2. Design of signalling system using a “what-if” approach for evaluating intervention scenarios	177
5.3. Designing an equi-block signalling layout maximizing economic efficiency of investment costs for a MRT line.....	180
5.4. Identifying the equi-block signalling layout which guarantees the best “trade-off” between users’ satisfaction and investment costs for a MRT line.	187
5.5. A model for supporting RAM analysis: towards a hybrid integration with an “event-driven” mesoscopic model.....	192
<i>5.5.1. Mesoscopic Model.....</i>	<i>193</i>
<i>5.5.2. Quantification of differences between the two models in terms of results accuracy and computational efficiency.</i>	<i>195</i>
Quantification of differences between the two models in terms of results accuracy.....	197
Quantification of differences between the two models in terms of computational efficiency	199
<i>5.5.3. Identification of the dynamic integration strategy between the two models.</i>	<i>200</i>
5.6. A model for evaluating effects on both Service Availability and Quality of Service: integration with a module for simulating passenger demand.	201
<i>5.6.1. Application to a MRT line</i>	<i>203</i>
Failure scenario 1: Train performance reduction of 60%.....	206
Failure scenario 2: Train performance reduction of 25%.....	208

5.6.2. <i>Towards an integrated simulation framework for Service Availability and Quality of Service evaluation to support RAM analysis</i>	209
Chapter 6. Conclusions	213
References.....	220

Chapter 1. Introduction

1.1. Motivations and Objectives of research

The objective of the work presented in this thesis has regarded the development of a “multi-purpose” microscopic infrastructure model for simulating railway systems, with the intent to overcome applicability limits of commercial microscopic models and support different design activities.

Due to the large amount of input data considered, microscopic infrastructure models available on the market are in fact inefficient to simulate large-sized networks or for being employed into analyses which require a large number of simulations (e.g. probabilistic analyses, “black-box” optimization) or short computing times (e.g. real-time applications). In these cases users are forced to rely on less detailed (e.g. macroscopic, mesoscopic models) or “fixed-speed” (e.g. based on alternative graphs) simulation models which are more efficient but inaccurate in results especially when high congestion levels are on the network.

Moreover, the closed-architecture which characterizes commercial microscopic infrastructure models, does not allow them to be used in applications where the interfacing with automatic mathematical structures (e.g. “black-box” optimization algorithms) or external softwares (e.g. for communicating filed data during real-time applications) is needed. For this reason in fact, practitioners need to have recourse to the development of customized models, which are often devoid of general validity (e.g. applicable only to the case-study examined) and respond only to their specific necessities.

The need felt by railway designers to handle more and more accurate, flexible and efficient microscopic models which effectively can support them during the different design activities, has raised the question of relying on appropriate simulation tools that satisfy such necessities, overcoming criticalities of current models.

To this aim, the research activity described in this thesis has concerned with the development of a microscopic infrastructure model, that can support designers in different application contexts, assuring contemporarily: *results accuracy* (since it is a high-detailed object-oriented model), *flexibility* (an open-structure allows the interfacing

with external softwares) and *efficiency* (a parallel architecture reduces computing times on multi-cores computers).

1.2. Thesis description and contributions

The work presented in this thesis can be slit up into two main phases: the first phase (*Development phase*) has regarded model specification and its practical implementation, while the second phase (*Application phase*) has dealt with applications of the realized model for solving several design problems.

The first step of the *Development phase* was relative to the definition of model features as well as its specification. The main features that the model had to have, strongly conditioned the choice of its structure, the architecture and the implementation technique. In fact the model has been developed in C++ by using an “object-oriented” technique in order to consent a detailed description of characteristics and dynamics relative to both infrastructure components (e.g. rail tracks, signalling equipment, stations) and rail vehicles (e.g. weight, length, number of wagons, “tractive effort-speed” curve of traction units, etc.). Furthermore it has an open-structure that allows the easy interfacing with external applications or automatic structures to perform for instance “black-box” optimizations or probabilistic analyses (e.g. sensitivity analysis). The architecture of the model is composed of the following four main modules whose respective input data can be directly set by the user through external files (containing data relative to the case-study under investigation):

- An *infrastructure module*, in which network is represented as a “link-oriented” graph model, where nodes contain information about positions of point elements (stations, block section joints, junctions, etc.) while links contain all rail track attributes (e.g. lengths, gradients, radii, speed limits, etc.).
- A *rolling stock module*, in which all physical and mechanical rail vehicle features are here considered as inputs. Therefore lengths, the number of coaches, wagon weights, as well as the “tractive effort-speed” curve of the traction unit, all must be specified to such module. Train movement on the track is simulated by integrating the Newton’s motion formula taking into account both track and vehicle motion resistances.

- A *signalling system module*, describes signalling equipments as a “discrete-event” model, in which signal aspects are changed according to the occupation state of block sections that they protect. Three different signalling technologies have been implemented (multi-aspect system, ETCS level 1 and ETCS level 2). Therefore the type of technology as well as the block section layout (e.g. their lengths and positions) constitute input data to this module.
- A *Timetable module*, requires that all information about train arrival/departure times at/from each station must be set as input. Moreover, it is also possible to specify train dwell times at stations. However such variables can be considered as deterministic or random variables distributed according to a certain probability function.

Such modular architecture assures the general validity of the model since it consents the application to whatever case-study, by simply specifying for each module the respective input data.

Successively the architecture of the model developed has been parallelized in order to decrease its computing performances when running on multi-core computers. Computing tests have in fact showed that benefits due to the parallelization on computing times increase when increasing the number of cores and dimensions of the problem, and therefore consent the efficient employment for large-sized networks or probabilistic analyses.

A validation phase has been then carried out. In particular a first verification phase of the code consisted in verifying the congruence of the output returned by the model with those given by a consolidated model (*OpenTrack*) for the same input dataset. Successively a validation of the model has been realized comparing simulated train performances with those observed on a real Mass Rapid Transit line: the “Cumana” line in Naples city (Italy).

A contribution to the analysis of model uncertainty has been possible, thanks to the features of the model developed, which have consented the interfacing with a mathematical framework for performing a sensitivity analysis. A sensitivity analysis has been in fact performed to understand how variability (or uncertainty) in model outputs depends on the variability of the different model inputs. The Sobol’ “variance-based”

technique has been applied, which allowed to investigate homogeneously the entire domain of input variables and estimate relative sensitivity indices (first-order and total indices). Results obtained considering the real case-study of the “Cumana” line have highlighted for certain network performances (model outputs), the design variables (model inputs) on which an intervention is necessary to efficiently improve them. The importance of performing a sensitivity analysis to support design stages has been therefore underlined, since it allows to identify which are the most relevant variables for a certain performance, and optimizing the allocation of economic resources intervening only on such variables and not also on the other parameters.

As said before, the second phase of the work developed in this research has regarded the application of the realized model to support several design activities.

An important step of this research work has consisted in integrating the microscopic infrastructure model within a “black-box” optimization loop for designing railway systems and/or operations according to a “what-to” design approach. The application of this design framework to the real case-study of the Cumana line, has consented to identify the optimal layout of an equi-block signalling system, which minimizes investment costs satisfying the level of capacity required by customers. Moreover, a further application has allowed to identify for a certain signalling technology, the equi-block section layout which offers the best trade-off between investment costs and user’s satisfaction.

An innovative modelling architecture has been defined for effectively supporting the so-called RAM (Reliability, Availability, Maintainability) analyses, which are probabilistic analyses requiring several millions of Monte-Carlo simulations, in order to verify if the *reliability* (inverse of the failure rate) and *maintainability* (probability to be maintained for a given operational condition and time period) of system components employed, as well as recovery strategies adopted, are able to satisfy service *availability* indices established by contract requirements. In particular such architecture will provide a dynamic integration with a mesoscopic model considering failure rates of components as input data. Such model is in fact very efficient since it is an event-driven “fixed-speed” model (based on the Stochastic Activity Network formalism), but its inaccuracy in results (especially for congested networks) can be compensated by the integration with the accurate microscopic model developed. An application of the two models to a

MRT line has consented to evaluate the trade-offs between computing efficiency of the mesoscopic and results accuracy of the microscopic. Outcomes have allowed to define the integration strategy of the two approaches: the microscopic model is launched a first time for calculating free-flow train running times which then are transferred as input data to the mesoscopic one; then computing efficiency of this latter model is exploited to perform millions Monte Carlo simulations of ordinary service, in order to draw (according to failure rates of components) stochastic failure events; only when a failure event is drawn the microscopic model is activated to accurately evaluate its effects on network performances.

Another fundamental part of this research work has regarded the interfacing of the microscopic model with a module for simulating effects on passenger demand (*demand assignment module*). This interfacing in fact has consented to assess impacts induced by a certain intervention not only on Service Availability (SA), but also on Quality of Service (QoS) delivered to passengers, as recently requested by standard European guidelines (given by the *CER*) for monitoring and measuring QoS on railway systems. The application of such simulation framework to a MRT line has shown that the most efficient solution in terms of network performances (SA) not always guarantees the highest QoS perceived by passengers, and highlight the necessity of explicitly simulate the effects on both railway operations and passenger travel demand.

Then, results obtained from the two previous applications has allowed to define the modelling strategy for a more complex simulation architecture which dynamically integrates the microscopic and the mesoscopic model, as well as the demand assignment module. Such architecture in fact, will consent to effectively support different kinds of design applications, returning impacts of interventions, disturbances or breakdowns on both SA and QoS. Probabilistic analysis aiming at making inference on stochastic failure events and estimating their effects on SA and QoS can be also efficiently realized, in fact: millions Monte Carlo simulation of undisturbed ordinary service are performed by means of the mesoscopic model, while the microscopic one will be activated only when a failure is drawn and degraded operation are on the network to accurately estimate its effect on performances; then train outputs (returned by the mesoscopic model for ordinary conditions or by the microscopic for degraded conditions) will be first elaborated to calculate SA indices (e.g. punctuality, regularity) and then will be transferred to the demand assignment module as inputs; this latter in

turn, taking into account the hourly passenger demand flows, will assess passenger loads of train runs and effects on QoS delivered (e.g. in terms of the user's generalized cost).

1.3. Thesis Outline

The research activity presented in this thesis has been organized in five chapters.

In Chapter 2, a general overview on railway systems and their infrastructural and operational components is provided. A first part introduces time-driven and event-driven systems, as well as mathematical models generally used to describe them. Moreover a description of railway systems as hybrid continuous-discrete systems is given and its representation as a hybrid simulation model is explained. Then a second part depicts all components which constitute railway systems: the infrastructure, the rolling stock, computer-based control systems (e.g. signalling equipment and interlocking) as well as train operations represented by the timetable, recovery strategies, shunting movements, etc.

Chapter 3 instead gives an outlook on the state of the art of railway traffic simulation models and their application in practical fields. Their distinction is introduced on the basis of the detail level of network representation (macroscopic, mesoscopic and microscopic models), the approach (deterministic, stochastic) and the processing technique of events (synchronous, asynchronous). Then commercial or laboratory models belonging to each category are depicted and respective applications for solving practical design problems are illustrated.

In Chapter 4, regards the description of all activities through which the microscopic infrastructure model has been developed. First general features of the model and the programming techniques employed to meet each feature are described. Then the specification of the model is deeply depicted by showing variables, input data, functions and algorithms of each module that constitutes the model (i.e. infrastructure, rolling stock, signalling system and timetable modules). An illustration of outputs returned by the model is provided, and the validation phase carried out is accurately described. Furthermore the methodology applied for performing sensitivity analysis of the model as well as results obtained are reported, in the final section of this chapter.

Chapter 5 deals with the illustration of different model applications carried out to support several design activities. First an application is depicted which is addressed to evaluate two different infrastructural intervention through a classical “what-if” design approach. Second the integration of the microscopic model within a “black-box” optimization loop is illustrated and the application of such framework for solving two different problems relative to signalling design, are described. Third, after a brief depiction of features relative to the aforementioned “event-driven” mesoscopic model, an application of the two models to the same case study is shown, and differences between the two approaches are highlighted. Then a description of the integrated architecture is reported. Fourth, the interfacing with a demand assignment model is illustrated and results obtained from an application of such structure to a MRT line are commented. After that, the description of an integrated modelling architecture for the evaluation of both SA and QoS is provided.

Finally in Chapter 6 a brief summary of all activities developed in this research is first reported then conclusive considerations regarding the work presented are explained.

Chapter 2. A General Outlook on Railway systems: Infrastructure components and operations.

Planning and designing stages of both railway infrastructure components (e.g. rail tracks, station areas, signalling systems, etc.) and operational strategies (e.g. service timetable, special train movements to recover ordinary conditions after a service disruption, etc.) aim at satisfying economic and performance constraints as imposed by contract specifications as well as the overall foreseen transportation demand, assuring acceptable levels of service quality. The achievement of such objectives can be reached through accurate evaluations of effects induced on the network by different intervention alternatives, especially when stochastic disturbances (e.g. failures, extended train dwell times at stations, train conflicts, etc.) to ordinary service do occur. However the high complexity degree which characterizes interactions amongst the different components of railway system, generally prevents network behaviour to be described by closed-form analytical solutions and imposes to rely on apposite simulation techniques. Nowadays, the technological development in computer science and the advent of High Performance Computing, has eased the spread of virtual simulation in most fields regarding both industrial and academic research applications, and also in the case of railway areas, simulation models are used to support planners and designers at each level during decisional phases. According to the level of detail through which railway network is depicted, railway simulation models can be distinguished in *macroscopic*, *mesoscopic* and *microscopic*. Macroscopic models represent railway network at a high level of abstraction, considering only aggregated network information (e.g. stations' positions, lengths of inter-station tracks), without modelling station areas or signalling equipments. Due to their scarce accuracy in results, such models are usually employed within long-term planning tasks where only approximate information of infrastructure elements are available. In addition the low volume of input data required, let such models be computationally efficient for simulating large-sized networks. Mesoscopic models instead are constituted by a "multi-scale" structure which mixes portions of the network depicted at microscopic level (i.e. at a high level of detail) and portions, that considered as irrelevant for the overall investigation outcome, are modelled at a macroscopic level to minimize computational efforts. Thanks to this feature such models are particularly suitable to answer some strategic questions like the management of train movements on shunting yards or the implementation of operational strategies to

recover normal conditions after a breakdown. Microscopic models describe at high level of detail each component of railway infrastructure (e.g. lengths and gradients of rail tracks, speed limits, station areas, signalling system, physical-mechanical rail vehicles characteristics, etc.), allowing for the thorough modelling of components interactions and transient train dynamics in order to give back accurate train running times calculations also when stochastic disturbances arise in the network. Therefore, such models return precise estimations of effects on network performances caused by a certain intervention, but because of the large volume of input data involved, they lack in computational efficiency and are usually used to simulate medium-small sized networks.

On the basis of the approach instead, simulation models of railway networks can be branched in *deterministic* and *stochastic*. Deterministic models considers that the time instant within which a certain event occurs is a constant parameter. Therefore in this case all train arrival/departure times at/from stations as well as train running times are considered as deterministic parameters whose values tally with that established within the timetable. On the contrary, stochastic models assume that such time instants are random variables with a certain probability distribution function.

Moreover with respect to the processing technique, railway simulation models can be split into *synchronous* (or time-driven) and *asynchronous* (or event-driven) models. In particular for the former, the simulation process goes ahead according to a clock-time, simulating the state of each component at each time-step. For the latter kind of models instead, the state of network components are updated only in correspondence to the realization of discrete events which occur in the network following a list of time stamps that set processing time instants (or the average processing time in case of stochastic models) for each event.

However simulation models of railway traffic are largely employed to different purposes in supporting both off-line and on-line decisional activities. In particular off-line applications essentially regard planning and designing phases of infrastructural or operational interventions, or stability analyses of service timetables, while on-line applications are mainly aimed at giving aid to railway dispatchers and train operators in efficiently rescheduling train runs or in managing recovery operations when disruptions occur during service. In addition recent on-line applications can be observed in train on-

board installations addressed to suggest train-drivers in order to perform “energy-efficient” driving strategies, according to the actual conditions of both the train itself and the surrounding traffic on the network.

In this chapter after a brief introduction to systems and models characteristics, a detailed description of the different categories of railway simulation models will be realized. Then a wider review of their applications in different railway fields is carried out.

2.1. Introduction to Time-driven and Event-driven systems

Generally a system can be defined as a physical entity composed of several interacting or interdependent components, forming an “integrated whole” which in turn reacts to determined external actions producing certain reactions (*Di Febbraro, Giua 2002*). Moreover system reactions (also called as effects) continuously evolve during time in order to fit to dynamic changes of external solicitations (also called as external causes). Dynamic evolution of both external actions and system reactions can be monitored by measuring the value of physical parameters which respectively represent them. In particular parameters which characterize external actions are called as *input* parameters and their time evolution is usually independent on system characteristics. Parameters relative to system reactions are instead known as *output* parameters, and their dynamic variation depends on both system features and external solicitations (therefore on input parameters). Substantially a system can be conceived as an operator with a certain complexity degree (often very high) which transforms a certain m -dimensional vector of input parameters $u \in R^m$ into a determined p -dimensional vector of output parameters $y \in R^p$. Furthermore the determination of the vector of system outputs $y(\tau)$ at a certain time instant τ , not only depends on the vector of input parameters at that instant $u(\tau)$, but also on system conditions within previous instants (i.e. on $y(\tau - t)$, $0 < t \leq \tau - 1$). Therefore in order to express an exhaustive input-output relationship which takes into account also for both current and previous system state conditions, a third n -dimensional vector of parameters $x \in R^n$ must be considered, whose elements represent “state variables” of the system. The value assumed by such a vector at instant τ , $x(\tau) = (x_1(\tau), x_2(\tau), \dots, x_n(\tau))$, depicts the state of the system at that instant having also considered the dynamic history of the system until that time, namely the values of system’s state variables within previous instants, $x(\tau - t)$, $0 < t \leq \tau - 1$. The definition of state variables vector consent therefore to express analytically the input-output

relationships which describe the dynamic evolution of system's behaviour. In literature such relationships are better known as “state equations” and according to the kind of system considered can be represented as:

- Time continuous differential equations for dynamic *continuous-time* systems:

$$\begin{cases} \frac{\partial x(\tau)}{\partial \tau} = f(x(\tau), u(\tau), \tau) \\ y(\tau) = g(x(\tau), u(\tau), \tau) \end{cases} \quad (1a)$$

- Time discrete chained equations for dynamic *discrete-time* systems:

$$\begin{cases} x(\tau) = f(x(\tau-1), u(\tau-1), \tau-1) \\ y(\tau) = g(x(\tau), u(\tau), \tau) \end{cases} \quad \tau = 0, 1, 2, \dots, N \quad (1b)$$

- System of equations of the type shown in (2) for *discrete-event* systems:

$$x(k) = \delta(x(k-1), e_k), \quad k = 1, \dots, K \quad (2)$$

where δ represents the so called “state transition” function, while e_k is the k^{th} event.

Equations (1a) and (1b) constitute state equations commonly employed in the field of classical physics for describing “*time-driven*” systems, whose state evolves with time, while equation (2) typically depicts the behaviour of “*event-driven*” systems whose state variables values change in correspondence to the occurrence of a certain event. Actually *hybrid* systems composed of both time-driven and event-driven components do exist, and their behaviour is therefore analytically described by very complex sets of equations containing both the first ((1a) and (1b)) and the second type (2) of relationships.

In order to perform studies and analyses of such systems, it is necessary to build opportune models of the system under consideration. In particular a *model* is a mathematical representation of the system, which consents to reproduce the behaviour of the system itself and the interactions amongst its components. Formal models addressed to depict time-driven systems are usually constituted by sets of time continuous or time discrete differential equations as the kind illustrated in (1) to determine time variations of both system's state variables ($x(\tau)$) and output parameters ($y(\tau)$) according to input data ($u(\tau)$) to the system. Such kind of models are generally called as Time-Driven Models (TDM). Models used to describe event-driven systems are instead known as Event-Driven models (EDM) and are constituted of sets of finite equations like the type shown in (2) that consider the whole set of events' traces,

namely all possible chronological sequences of events that can be generated by a discrete-event system. In particular according to the level of abstraction through which are represented the sequences of possible events (traces), EDM can be split into: Logical Event-Driven Models and Timed Event-Driven Models. The former consider events' traces as a list of events (e_1, e_2, \dots, e_K) ordered according to their sequence of occurrence, but without giving any information about the instant in which they arise. Therefore such kind of models can return only information about the ordered k -dimensional vectors of both system's state (x_1, x_2, \dots, x_K) and output parameters (y_1, y_2, \dots, y_k) corresponding to each i^{th} event $(i \in 1 \dots k)$, but nothing can be determined about time instants in which such state transitions do occur. Within a Timed Event-Driven Model instead, events' traces are represented as an ordered k -dimensional vector of couples $((e_1, \tau_1), (e_2, \tau_2), \dots, (e_K, \tau_K))$ containing information about both the i^{th} event e_i and the respective instant of occurrence τ_i . As a consequence, for this type of models is possible to have also information on time instants in which state transitions x_i arise. Moreover, Timed EDM can be in turn branched into *deterministic* or *stochastic* models according to assumptions made on time instants τ_i associated to the i^{th} event e_i . Deterministic models in fact assume that elements of the vector $(\tau_1, \tau_2, \dots, \tau_k)$ also called as "clock structure", are deterministic variables (i.e. they have constant values). On the contrary stochastic models consider that time instants of such a vector, are random variables with a known probability density function. Therefore for stochastic models the "clock structure" vector tallies with the k -dimensional vector composed of probability density functions (pdf) $(\varphi_1, \varphi_2, \dots, \varphi_K)$ relative to each time variable τ_i , and for this reason it is called as "stochastic clock structure". Anyway since Timed EDMs take into account for the instant of occurrence relative to each event, it is possible to determine also the time trace of system's states, namely the sequence of time instants $((x_1, \tau_1), (x_2, \tau_2), \dots, (x_K, \tau_K))$ in which the i^{th} state transition x_i does arise.

Generally Logical EDMs are usefully employed to study qualitative characteristics of an event driven system therefore allowing for a structural analysis, while Timed EDMs are mostly used when realizing performance analyses of the system, since they can also return the chronological sequence in which both events and state transitions occur.

2.2. Railway networks as hybrid continuous-discrete systems

Railway networks are very complex systems which are composed of many components (Figure 1) that can be mainly grouped into:

- *Infrastructure* (e.g. stations, rail tracks, etc.)
- *Timetable and operation strategies*
- *Rolling stock* (e.g. rail vehicles)
- *Signalling equipments and other computer-based systems* (e.g. interlocking systems, level crossings)

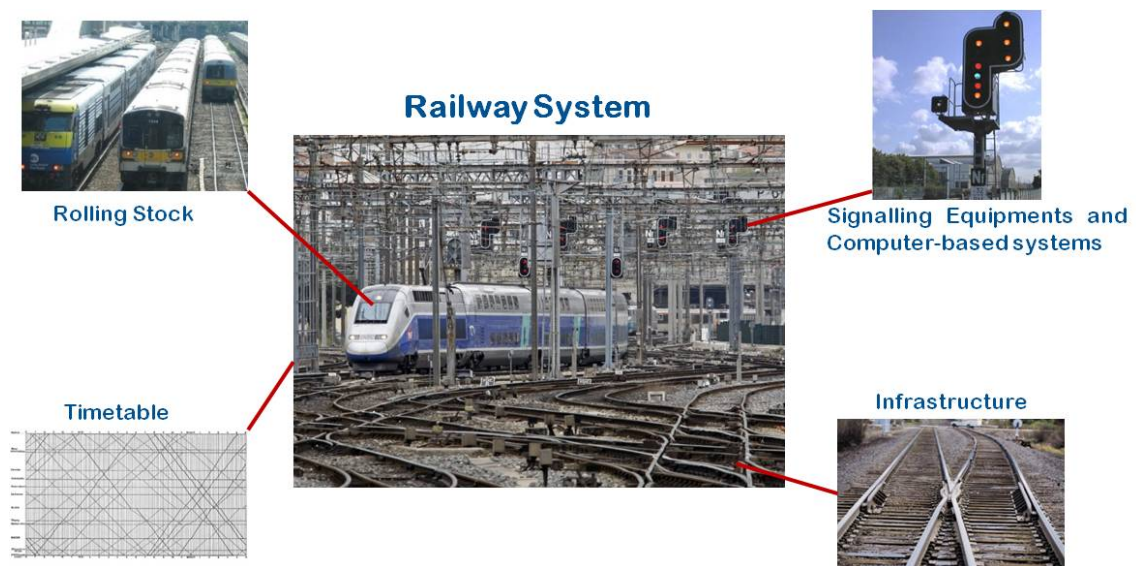


Figure 1. Railway system and its components

Such components strongly interact during railway service and the dynamic evolution of these interactions determine the configuration of system dynamics during time. In particular the component “*Infrastructure*” involves all physical elements of railway network, like for example rail tracks, station platforms, pocket tracks to store away corrupted trains, power substations areas, and each subcomponent which guides train movements along space dimension. Instead the component “*Timetable and operation strategies*” include scheduled train arrivals/departures at/from stations, scheduled dwell times, scheduled train running times, as well as all the other planned time points (e.g. periodic shunting movements for vehicle maintenance) which regulates train movements within time dimension. It is clear that the state of these two components of railway system does not change frequently during time, since they constitute the basic

space-time structures on which railway service is managed. Moreover their characteristics derive from long and complex planning and designing phases in which both infrastructural and operational solutions are determined in order to satisfy both transportation demand and environmental objectives, respecting technical and technological constraints imposed by both contracts requirements as well as geomorphological conformations of the surrounding sites. Indeed, variations in the layout of infrastructural elements such as rail tracks or station platforms can be observed only few times during the entire system lifecycle, for example in correspondence to renewal works, extraordinary maintenance tasks, or reparation works of corrupted items. On the contrary, variations of timetables and other scheduled operations can occur more frequently during system lifecycle, since if a change in system performances is required by surrounding conditions (e.g. increase in demand level), acting on timetable is surely cheaper and easier than intervening on infrastructural elements. Actually train operations can be rescheduled by dispatchers, many times also within the same working day, in order to efficiently recover ordinary service after operation disturbances induced by unforeseen events like for example train delays or breakdowns. In fact “dispatching centres” continuously monitor traffic conditions in real-time, deciding to apply a reorganization of train runs (rescheduling) or a particular operational strategy (e.g. storing away a broken train on a pocket track) when an anomaly in system working conditions such as failures or strong delays do occur in the network. Obviously, the intense decisional activity developed by railway dispatchers in managing large volumes of railway traffic is effectively supported by apposite simulation tools which can predict within short times the effects on network performances induced by a certain operational solution taking into account current traffic conditions, consenting therefore to identify the best intervention alternative.

“*Rolling Stock*” component involves rail vehicles and their movement on the track. In particular rail vehicles are often composed by a traction unit (electrical or diesel unit) supplying the tractive effort necessary to pull the coaches (containing passengers and /or freight) and move the train. With respect to the previous two components examined above, rail vehicles can be instead considered as continuous-time subsystems since their state parameters (e.g. spatial position, speed, acceleration) continuously evolve during time when service is on. For this reason their behaviour can be described by using a *Time-Driven* model regulated by equations of the type illustrated at (1a) and (1b).

“Signalling equipments and other computer-based systems”, are usually considered as a part of the infrastructure component, also if their role is quite different with respect to train movement. In fact such subsystems regulates train movement on the track in order to keep always safe conditions on the network both in the open track (railway sections outside station areas) and within station or shunting areas. In particular signalling systems installed on the track are addressed to perform the principle of train separation, checking with a certain frequency (which depends on the technology used) the positions of two consecutive trains and regulating the movement of the follower train in order to keep an acceptable safety margin with respect to the leader train. For example if line-side signals are installed on the track and a certain positions is occupied by train A, they will give a red aspect to the follower train when this one will be at a certain safety distance (usually imposed by the block section length) by train A. Anyway other computer based systems such as Interlocking centres (ACCM, ACEI, etc.) and level crossings, respectively regulate train movements within station areas and intersections with other infrastructures like roads. In particular Interlocking systems are mainly addressed to perform feasibility checks for safe train entrances within station areas according to the surrounding traffic conditions and set up station route for entering trains, in order to keep safe conditions in the station areas and let trains reach platforms for stopping operations (e.g. passengers alighting/boarding, exchange of the on-board crew, waiting for a connecting train). As can be seen, signalling equipments and all the other computer-based systems can be certainly classified as discrete-event systems, since their state conditions change only in correspondence of a discrete event (e.g. a line-side signal change its aspect only when a train occupies a certain position), therefore the behaviour of such items can be described by an Event-Driven model controlled by equations of the type reported at (2).

At this point it is clear that since railway systems are composed by both Time-Driven (rail vehicles) and Event-Driven systems (computer-based equipments), it can be classified as a Hybrid continuous-discrete system and therefore its behaviour is described by a Hybrid continuous-discrete model regulated by complex systems of equations which involve both equations of the kind reported at (1a), (1b) and (2).

Actually the high degree of complexity which characterizes interactions amongst network components, prevents a closed-form solution for the aforementioned equation system, and therefore complex dynamics of railway system can only be described by

means of apposite simulation techniques. Models for simulating railway networks explicitly describe the dynamics of each component of the system as well as their respective interactions. However to better understand the role and the working modes of each elements with respect to railway system dynamics, a deeper description of each one of the components listed above is necessary.

2.3. Railway infrastructure

As said in the previous section, railway infrastructure includes both rail tracks and stations. In particular tracks consist of two parallel steel rails, anchored perpendicular to members called ties (sleepers) of timber, concrete, steel, or plastic to maintain a consistent distance apart, or gauge. The track guides the conical, flanged wheels, keeping the cars on the track without active steering and therefore allowing trains to be much longer than road vehicles. The rails and ties are usually placed on a foundation made of compressed earth on top of which is placed a bed of ballast to distribute the load from the ties and to prevent the track from buckling as the ground settles over time under the weight of the vehicles passing above (Figure 2).

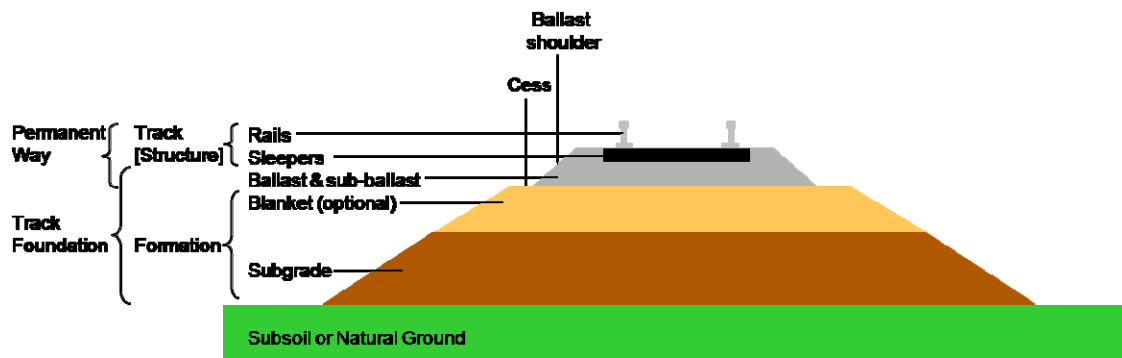


Figure 2. Layout of Railway Infrastructure foundation and permanent way

The ballast also serves as a means of drainage. Some more modern track in special areas is attached by direct fixation without ballast. Track may be prefabricated or assembled in place. By welding rails together to form lengths of continuous welded rail, additional wear and tear on rolling stock caused by the small surface gap at the joints between rails can be counteracted; this also makes for a quieter ride (passenger trains). Spikes in wooden ties can loosen over time, but split and rotten ties may be individually replaced with new wooden ties or concrete substitutes. Concrete ties can also develop cracks or splits, and can also be replaced individually. Should the rails settle due to soil subsidence, they can be lifted by specialized machinery and additional ballast tamped under the ties to level the rails.

A great attention must be paid, for curved rail tracks since on this elements train wheels have both a pure rolling movement and a certain rubbing against the rails, which determines the arising of additional vehicle resistances, and overall imposes speed limits (varying proportionally with the curve radius) that train runs must respect to prevent dangerous vehicle derailments. In particular on curved sections the outer rail may be at a higher level than the inner rail. This is called super-elevation or cant. This reduces the forces tending to displace the track and makes for a more comfortable ride for standing livestock and standing or seated passengers. A given amount of super-elevation will be the most effective over a limited range of speeds. In particular the maximum amount of super-elevation h adopted by the Italian Infrastructure Manager “RFI” is 0.16 m. To calculate the speed limit corresponding to a certain curve radius ρ , it is necessary to consider the vehicle which is running on a curved rail section with a gauge s , as depicted in Figure 3.

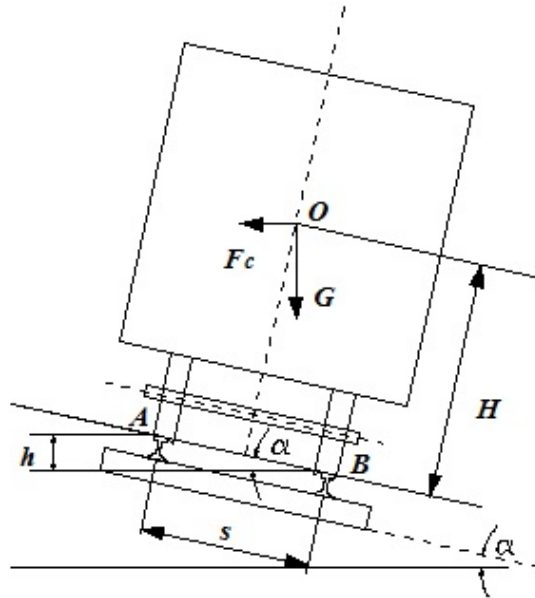


Figure 3. Scheme of a rail vehicle running on a curved section

Let be $\alpha = \arcsin(h/s)$, it is possible to express the vertical component G and the horizontal component F_c' , of the centrifugal force F_c as:

$$G = m \cdot g \cdot \sin \alpha, \quad F_c' = \frac{m \cdot v^2 \cdot \cos \alpha}{\rho}; \quad (3)$$

Where m is the vehicle mass, while g represents the gravitational acceleration. As a consequence the so called *non-compensated centrifugal force* F_{nc} to which both the vehicle and passengers are subjected to is therefore calculated as:

$$F_{nc} = F_c - G = \frac{m \cdot v^2 \cdot \cos \alpha}{\rho} - m \cdot g \cdot \sin \alpha ; \quad (4)$$

Since $\cos \alpha \approx 1$, given that α values are near to zero it is possible to express the *non-compensated acceleration* a_{nc} as:

$$a_{nc} = F_{nc} / m = \frac{v^2}{\rho} - g \cdot \sin \alpha = \frac{v^2}{\rho} - g \cdot \frac{h}{s} ; \quad (5)$$

Generally the value of a_{nc} to which passengers are subjected to does not have to overcome 1 m/s^2 . Anyway from relation (5) it is immediate to obtain the speed limit v_{lim} corresponding to a curve with radius ρ , as:

$$v_{lim} = \sqrt{a_{nc} + g \cdot h / s} \cdot \sqrt{\rho} ; \quad (6)$$

Actually such speed limits are usually different according to the category of train service supplied (e.g. metro, intercity, regional, etc.), in fact for each one of this categories a different value of a_{nc} is considered. For example the Italian “RFI” for a certain curved section usually establishes three ranks of speed limits respectively characterized by the following values of *non-compensated acceleration* a_{nc} :

A) $a_{nc} = 0.6 \text{ m/s}^2$;

B) $a_{nc} = 0.8 \text{ m/s}^2$;

C) $a_{nc} = 1.0 \text{ m/s}^2$;

The other fundamental element composing railway infrastructure is constituted by stations, which can be classified as punctual structures where the so called “train-stopping” operations are developed like for example loading/unloading of passengers and/or goods, exchange of on-board staff, etc. Stations generally consist of a platform next to the tracks and a station building providing related services such as ticket sales and waiting rooms. If a station is on a single track main line, it usually

has a passing loop to facilitate the traffic. It is possible to make a classification of train stations according to their own layout in terms of their physical extension, the number of platforms, the complexity of station tracks and their position within the network. In particular, it is possible to distinguish among:

2.3.1. Terminal Stations

A "terminal" or "terminus" is a station at the end of a railway line (Figure 4 a). Trains arriving there have to end their journeys (terminate) or reverse out of the station. Depending on the layout of the station, this usually permits travellers to reach all the platforms without the need to cross any tracks – the public entrance to the station and the main reception facilities being at the far end of the platforms. Sometimes, however, the railway line continues for a short distance beyond the station, and terminating trains continue forwards after depositing their passengers, before either proceeding to sidings or reversing back to the station to pick up departing passengers. A terminus is frequently, but not always, the final destination of trains arriving at the station. However a number of cities, especially in continental Europe, have a terminus as their main railway stations, and all main lines converge on this station. There may also be a bypass line, used by freight trains that do not need to stop at the main station. In such cases all trains passing through that main station must leave in the reverse direction from that of their arrival. There are several ways in which this can be accomplished:

- arranging for the service to be provided by a multiple unit, or push-pull train, both of which are capable of operating in either direction. The driver simply walks to the other end of the train and takes control from the other cab. This is increasingly the normal method in Europe.
- by detaching the locomotive which brought the train into the station and then either
- using another track to "run it around" to the other end of the train, to which it then re-attaches;
- attaching a second locomotive to the outbound end of the train; or
- by the use of a "wye", a roughly triangular arrangement of track and switches (points) where a train can reverse direction and back into the terminal (Figure 4 b).



Figure 4. a) Example of train terminal station: Hamburg Hauptbahnhof, b) Usage of a “wye” track for turning a rail vehicle

Some former termini have a newer set of through platforms underneath (or above, or alongside) the terminal platforms on the main level. They are used by a cross-city extension of the main line, often for *commuter trains*, while the terminal platforms may serve long-distance services.

2.3.2. Ordinary Stations

Such elements are usually distinguished by Terminal Stations, since they are not located at the end of the network, neither they are equipped with a depot to store rail vehicles. However, these infrastructures can be composed of different platforms for performing train stopping operations, and beyond the main track can be equipped with further tracks, usually called “sidings” where it is possible to make up, assemble or store away trains (Figure 5 a). Anyway, sidings cannot be used for regular train movements but only for *shunting* movements, i.e. movements performed at lower speeds (25-30 km/h) allowed by means of particular train movement authorizations. Train routes within ordinary stations are usually controlled by interlocking systems which automatically authorize train movements within station areas, and provide the right positions for points and switches to let trains respect their own scheduled route. According to the kind of interlocking system implemented, stations can be monitored by an operator directly placed within the station area or may be operated via remote or distant control, by a Centralized Traffic Control (CTC). Stations may also be classified according to the layout of the platforms. Apart from single-track lines, the most basic arrangement is a pair of railway tracks for the two directions; there is then a basic choice of an island

platform between, or two separate platforms outside, the tracks. With more tracks, the possibilities expand.

Some stations have unusual platform layouts due to space constraints of the station location, or the alignment of the railway lines. Examples include staggered platforms, such as at Tutbury and Hatton railway station on the Derby-Crewe line, and curved platforms, such as Cheadle Hulme railway station on the Macclesfield to Manchester Line. Triangular stations also exist where two lines form a three-way junction and platforms are built on all three sides (Figure 5 b).

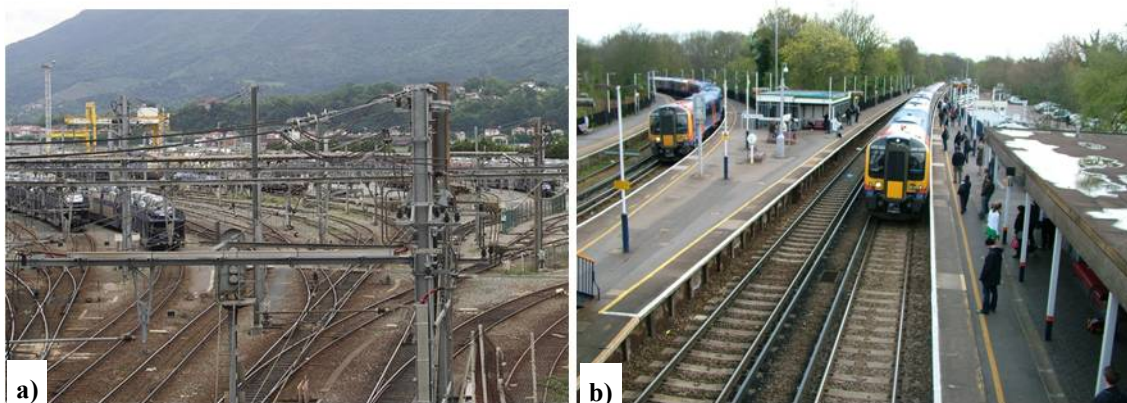


Figure 5. a) Sidings at Hendaye Station (France), b) Triangular platform at Virginia Water Station (UK).

2.3.3. Stop Stations

A railway stop is a spot along a railway line, usually between stations or at a seldom-used station, where passengers can board and exit the train. While a junction or interlocking usually divides two or more railway lines or routes, and thus has remotely or locally operated signals, a station stop does not. A station stop usually does not have any tracks other than the main tracks, and may or may not have switches (points, crossovers). The role of stop stations is therefore merely addressed to consent passengers boarding or alighting from a railway line, and given that their platform are located directly on the main track where regular train movements are performed, they usually do not require the installation of apposite interlocking systems for the management of train routes and switches. For this reason they are generally not assisted by any local operator. For example Figure 6 illustrates the “Artarmon” stop station in Sidney, where it is clear that no tracks other than the main

tracks are present, and that no switches, points or interlocking systems are necessary since platforms directly serve main tracks.



Figure 6. Artarmon stop station in Sidney

Moreover along the open track it is possible to find other kind of sidings, usually called as “pocket track” (Figure 7) whose main role concerns with putting in practice particular operational strategies addressed to recover normal service after a disruption due for example to a breakdown of a rail vehicle. Moreover they can be also used for turning train and reverse their direction. However when a rail vehicle is subjected to a failure during service, this will be usually stored away on the nearest pocket track, in order to recover ordinary as soon as possible. Therefore pocket tracks give the possibility to rapidly put away broken vehicles, and decide whether repairing the vehicle on site (if possible) and put again the repaired train on service starting from there, or waiting for the end of service to bring back the corrupted train to the depot where it will be subjected to necessary fixing procedures.



Figure 7. Pocket Track on the Washington metro

2.4. Timetable

Timetable is the result of complex planning and designing processes through which all train operations are scheduled in order to satisfy the demand level loading the network, assuring a certain robustness and stability towards stochastic disturbances that can arise during real service tasks. Therefore, timetable dictates working conditions of the entire railway network, establishing train dwell times at stations, arrival or departure times from platforms, connections between runs, train headways, but also operations which are not directly related to ordinary service, such as shunting movements on siding tracks for composing or maintaining rail vehicles. To design an effective service timetable, it is necessary to start from train running times on the network in order to understand the times needed by each train to complete its path, but overall to identify critical network sections, such as bottlenecks or singular sections when a certain kind of train conflicts can arise. In particular the scheduled running time of a train consists of :

- Pure running time between scheduled stops
- Dwell time at scheduled stops
- Recovery time
- Scheduled waiting time.

The pure running time between scheduled stops is the shortest possible running time as a result from a running time calculation (*Hansen and Pachtl, 2008*). Moreover, it is necessary that the pure train running time must be increased of a certain amount of recovery times to allow a train to make up small delays. In particular two different types of recovery time can be listed:

- Regular recovery time
- Special recovery time.

The “*regular recovery*” time is added to each train run as a percentage of the pure running time. Generally their values are in the range of 3-7% within European area and 6-8% in North American countries for passenger trains. Such time supplements, can be moreover spread over the whole train path or concentrated at singular points which are frequently identified as sources of delay, such as large intermediate terminals or congested network joints or stations (Figure 8).

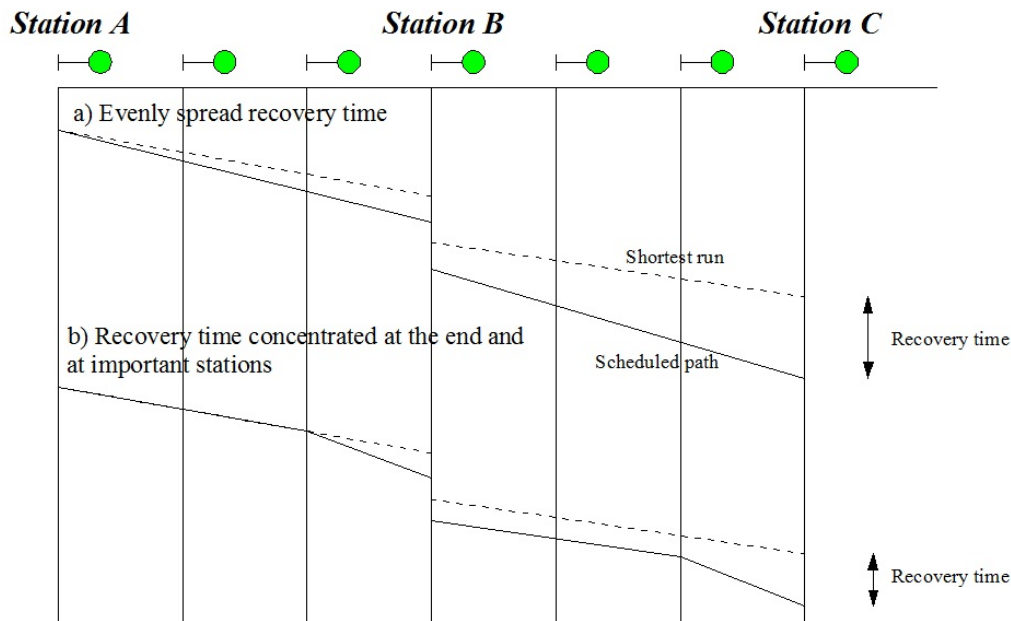


Figure 8. Distribution of recovery time for timetable construction: a) evenly spread recovery time, b) recovery time located only at relevant nodes

The “*special recovery*” time is instead used to compensate for the influence of both maintenance works and sections with temporarily bad track conditions (*Hansen and Pachl, 2008*). Instead of the regular recovery time it is not added as a percentage of train running times, but it is added as a time supplement only at that sections concerning with the aforementioned kind of problems.

The “*Scheduled waiting*” time is instead added for scheduling reasons, e.g. for synchronizing different passenger lines at connecting stations, or waiting for scheduled passing or overtaking. Such kind of time supplement is usually added to scheduled dwell times at stations, but can be also added to pure running times.

2.4.1. Principles of Train separation

As said before, to design a timetable, a preliminary modelling of train paths is required in order to plan effective train run schedules which are free from any kind of conflicts. A train path describes the usage of the infrastructure for a train movement on a track and in time (*Pachl, 2002*). In general it is not sufficient to describe train paths by means of their train run graph, but it is necessary to consider the whole “time channel” that train movements produce during their space-time line. Such issue has been faced and solved since the late 1950s by Happel, which developed a model to describe the time-channel related to a train path, the so called “*Blocking Time*” model (*Happel, 1959*).

Currently the blocking time model is used throughout both European and American countries for scheduling train operations of railway networks. A correct application of such model, requires anyway a basic knowledge of rules which regulate train movements during service, and in particular of train separation principles. As known, the adhesion coefficient between rails and rail vehicle wheels is about eight times less than that observed within road systems, and for this reason braking distances of rail vehicle are generally longer than sight distances of train drivers, unless the train is performing shunting movements where maximum running speeds are lower than 25-30 km/h. Therefore during regular service, train movements and in particular decelerating phases to keep safe distances from preceding trains, need to be controlled by external systems (e.g. signalling equipments) which work independently from the viewing range of the driver. In particular the principle used for train separation depends on the following criteria:

- How movement authority is transmitted from track to trains
- How the line is released behind a train.

If movement authority is only transmitted at discrete points, e.g. fixed signals, or by written or verbal orders, this will lead to train separation in a fixed block distance since each movement has to cover the entire section up to the next point where further authority may be received (*Hansen and Pachel, 2008*). Such type of train separation is currently implemented on all railway lines where trains are governed by line-side signals and protected by intermittent Automatic Train Protection systems (e.g. coded track circuit, ETCS level 1). In particular to apply safe train separation rules the track must be subdivided into discrete portions called as “block sections”, which are defined as track sections that can be exclusively occupied by only one train. In a fixed block operation, therefore block sections are delimited by signals providing movement authority to enter the block section protected by the signals. A train can enter a block section only if the train ahead have cleared the block section (or the overlap behind the exit protection signal), and if its entering movement is protected from opposing train movements as well as following train movements, by apposite stop signals. Where train movements are instead governed continuously (indeed discrete systems with a high detection frequency like radio signalling equipments used in ETCS level 2) by a cab signal system, fixed line-side signals could be also omitted. However, the introduction

of continuous transmission of movement authority does not represent a sufficient criterion to eliminate fixed block sections. Moreover in these cases the train has to release the track not at fixed intervals but continuously. This requires a permanent train-borne checking of train integrity. Given that a solution to this kind of problem has not yet been found, train separation in fixed block distance is still the standard principle for safe train spacing on most railways worldwide. Anyway, when train separation is performed by means of a fixed block layout, the main advantage that cab signalling provide with respect to systems governed only by line-side signals, is the independence of the cab signals from the approach distance of the line-side signal system, i.e. the distance between the signal at the entrance of the block section and the signal in rear that provides the approach indication (*Hansen and Pachl, 2008*). This feature allows therefore trains to run at higher speeds and that is why cab signalling is today the standard system on all high speed railway lines. Actually on the most part of railway systems cab signalling is coupled with continuous ATP systems such as coded track circuits. Furthermore train separation rules not based on fixed block sections are nowadays under study to implement the so called “moving block” where the block section length is reduced to zero and train movements controlled by cab signalling. This is what the ETCS level 3 signalling system type proposes to realize, but as said before this system is currently not implementable due to the unsolved problem of train-borne checking of train integrity.

2.4.2. The Blocking Time model

The blocking time is the total elapsed time in which a section of track (e.g. a block section, an interlocked route) is allocated exclusively to a train movement and therefore blocked for other trains (*Hansen and Pachl, 2008*). Given such a definition, it is clear therefore that the blocking time for a certain section begins when a train starts requiring movement authorization to enter that section (which is usually given before the train has reached the braking distance in approach to that section) and ends only after the train has completely cleared that section and all signalling appliances have been reset to normal position so that movement authority can be issued by another train to enter the same section. It is clear that for this reason the blocking time is generally longer than the time in which the train actually occupies the section. Considering for example a track portion where no scheduled stops are involved and where train separation is

assured by fixed block sections with line-side signals, it is possible to compute the corresponding blocking time taking into account the following time intervals:

- The *time for clearing the signal*
- The *signal watching time*, i.e. the time needed by the driver to view the clear aspect at the signal that gives the approach indication (e.g. distant signals) to the signal at the entrance of the block section
- The *approach time* between the signal that provides the approach indication (e.g. distant signal) and the main signal at the entrance of the block section,
- The running *time between the block signals*
- The *clearing time*, namely the time required to unclear the block section and the eventual overlap with the full length of the train
- The *release time* to unlock the block system.

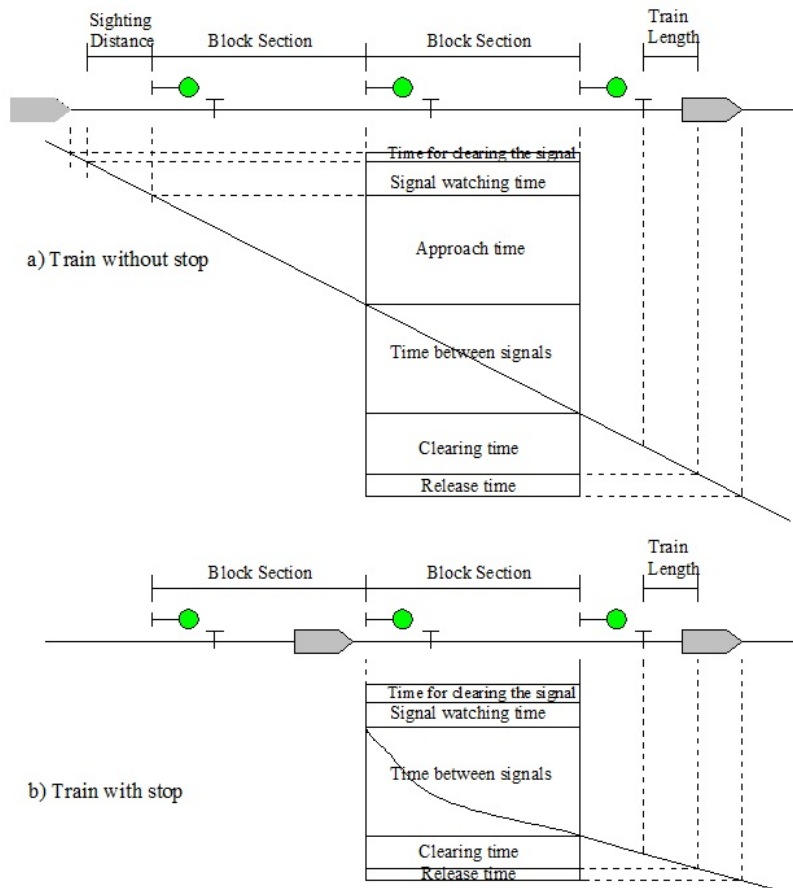


Figure 9. Computation of blocking times for a train running on a path: a) without stops and b) with stops

In Figure 9 it is represented how computing blocking time for a track section equipped with line-side signals both in absence (a) and in presence (b) of a scheduled stop.

When an ETCS level 1 signalling type is implemented on the track, the approach time is no longer the running time between the signal that provides the approach indication and the signal at the entrance of the block section, but the time the train runs through the braking distance that is supervised by the ETCS level 1 system. Since the braking distance calculated by such systems adopts a deceleration rate which is usually lower than those used within traditional ATP systems, this results in an increased approach time whose effects have been demonstrated by *Wendler (2006)*.

When drawing into a space-time diagram the blocking times for all block sections crossed by a train during its own path, it is possible to obtain the so-called “*blocking time stairway*”, which represents the operational use of a line by a train (Figure 10).

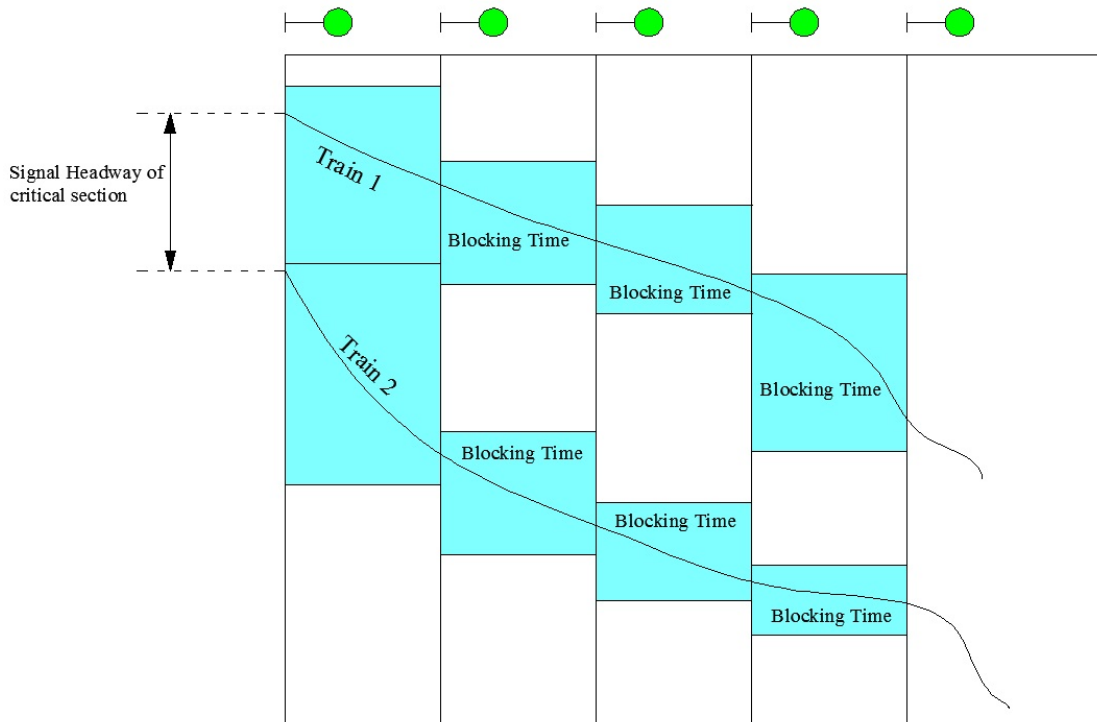


Figure 10. Blocking time stairways and Signal Headway for a simple railway line

In particular the use of blocking time stairways is fundamental to calculate the minimum headway that can be fixed between two trains, since it directly determines the so-called “*signal headway*”, namely the minimum time lag between two consecutive trains considering only one block section. When considering instead not only one block section but the blocking time stairways for the entire line between two stations, the maximum value amongst the signal headways represents the *minimum line headway*. In this case the blocking time stairways of two consecutive trains would touch each other without any tolerance in at least one block section called as “the critical block section”.

As illustrated in *Hansen and Pachl (2008)*, to calculate the minimum line headway of two trains, the train paths are virtually put one over the other imposing the same departure time in order to overlap for each block section the corresponding blocking times. Then for a certain block section the blocking time overlap equals the amount of time the train path of the second train must be postponed to eliminate the blocking time conflict in such a block section. (Figure 11).

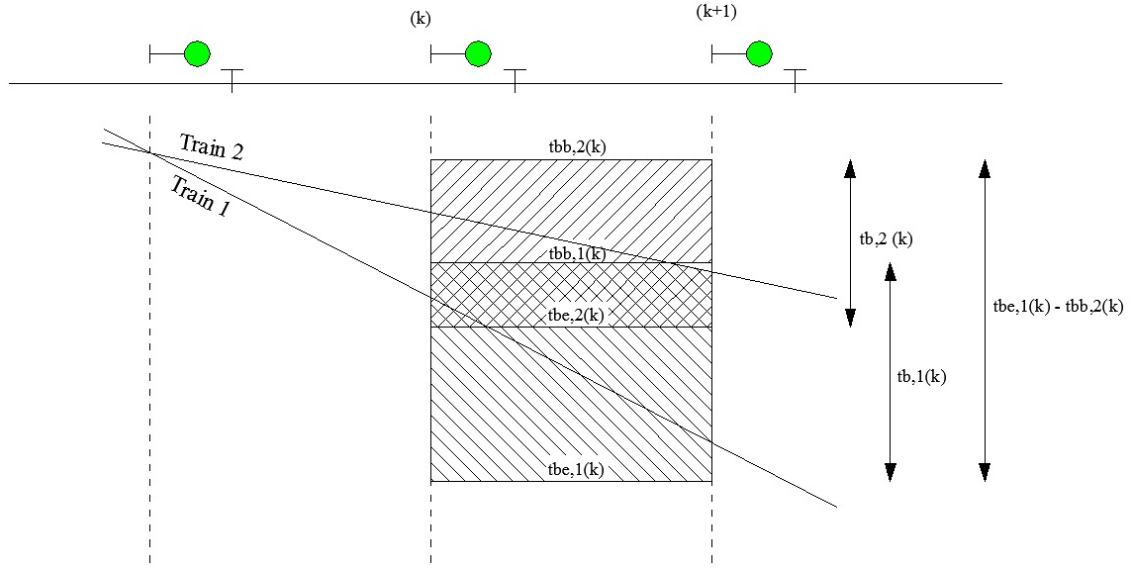


Figure 11. Calculation of Blocking Time overlap for a single block section

After calculating the amount of blocking time overlap (which corresponds to the signal headway for that section) for each one of the considered block sections, the maximum amongst such values tallies with the total time that the second train must be deferred with respect to the first train to eliminate all conflicts between the blocking time stairways, thus it represents just the minimum line headway. Therefore the minimum line headway $t_{h,ij}$ for train j following train i can be calculated as:

$$t_{h,ij} = \max \{t_{be,1(k)} - t_{bb,2(k)}\} \text{ for } k = 1 \dots n_b; \quad (7)$$

Where $t_{be,i(k)}$ is the end of the blocking time of train i in block section k , $t_{bb,j(k)}$ represents the beginning of blocking time of train j in block section k , while n_b is the total number of block sections considered.

However, to design a robust timetable, the time distance between two successive trains must be larger than the minimum line headway determined by means of the equation (7), since if the leading train is subjected to any kind of delay (large or small), also the following train will certainly experience a delay transferred from the first train (called as *knock-on* delay). Therefore to assure a certain level of robustness to scheduled train runs, the time distance between two consecutive trains must be equal to the minimum line headway increased by a certain buffer time to compensate for small delays. Hence, the buffer time is the smallest slot between the blocking time stairways of two trains. Anyway such time supplement must not be confused with the recovery time, since the

former prevents a small delay from being transmitted to other trains, while the latter enables a train to make up small delays which depend on its own run. Actually the amount of buffer times between trains depends on the quality of service required for a network, in fact the more is the quality of service needed, the larger are buffer times between consecutive train runs. It is clear that such buffer times cannot be greater of a certain amount or the adoption of very large buffer times will induce a reduction in system capacity.

Usually, in fact the buffer time is determined depending on the kind of line headway. Generally the rules adopted by the most part of railway systems are the ones listed below:

- Large buffer time when the second train has a higher priority than the first train
- Small buffer time when the first train has a higher priority than the second train
- Average buffer time when both trains have the same priority.

Indeed, the allocation of both recovery time to each single train run and buffer times between consecutive trains, strongly influence the stability and the robustness of the timetable itself, and for this reason many innovative methods are currently under examination for robust design of railway timetables. In particular procedures for a stochastic modelling of buffer time allocations has been implemented by Hansen (2004) and Kroon et al. (2007). Additionally, it is worth saying that buffer times are also required at connecting stations where scheduled transfer connections between trains exist and where crew or equipment changes from one train to another. To prevent the transmission of delays at such points, a apposite buffer time must be added to the time that is required for the transfer. Goverde (2005) developed a stochastic model of the transfer times to design effective buffer times at connecting points.

2.5. Rolling Stock

The basic unit of a railway is the wheelset (Figure 12 a). The conventional wheel set of today has the following features: it consists of two wheels fixed on a common axle, so that each wheel rotates with a common angular velocity and a constant distance between the two wheels is maintained. Flanges are provided on the inside edge of the treads and the flange-way clearance allows, typically 7-10 mm of lateral displacement to occur before flange contact. Whilst many wheelsets commence life with purely coned treads,

typically coned at 1/20 or 1/40, these treads wear rapidly in service, so that the treads come to possess curvature in the transverse direction. Similarly, rails also possess curvature in the transverse direction. All these features contribute to the behaviour of the railway vehicle as a dynamic system and it is important to consider their purpose.

According to the measure of track gauge (i.e. the distance between inner faces of rails) wheelsets may have different lateral extensions. The conical shape of rail wheels anyway contribute to ease rolling movements within curved rail sections, in fact for very large radii of curved tracks, the contact between the rail and the flange occur rarely. On the contrary, the flange of the wheel assumes an important role for sharp curves, where in addition to a pure rolling movement of the wheel on the track a strong contact between flanges and rails occur, which prevents the vehicle from dangerous derailments. A rail coach is composed of a “running gear” in turn constituted by several wheelsets (usually running gears with 2 or 3 wheelsets are common), while a rail vehicle is composed by one or more coaches which are moved by a traction unit, commonly called as “locomotive”.

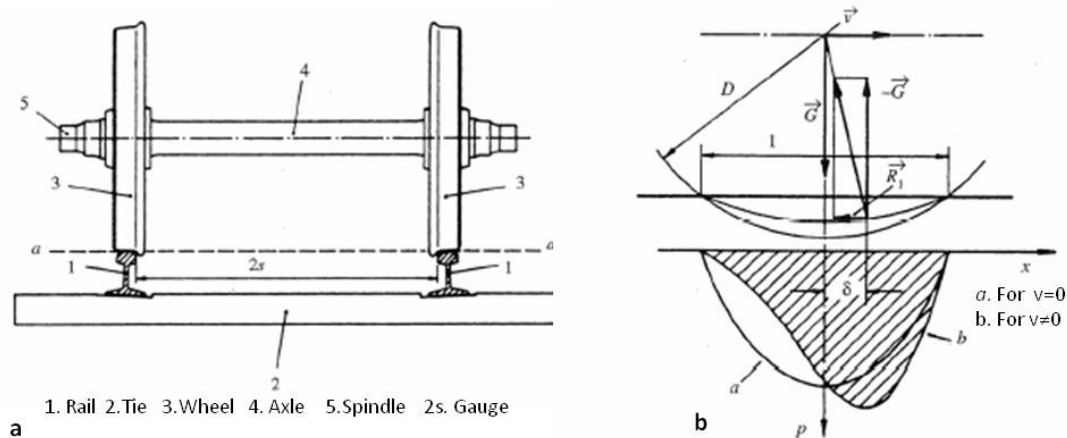


Figure 12. Scheme of a rail vehicle wheelset (a) and pression distribution on the contact area between wheel rim and the rail (b).

The contact between the wheel and the rails is concentrated in a finite area A , where the weight G deriving from the wheelset is distributed on (Figure 12 b), with a pressure

$$p = \frac{dG}{dA}, \text{ and therefore an average pressure value of } p_m = G / A \text{ which can reach values}$$

in the order of 500-1000 N/mm². When the wheel is at a standstill the contact area assumes an ellipsoidal or circular shape while if the wheel is fatigued, such area will assume a rectangular shape. However on this contact area pressures are distributed

according a parabolic law. Instead, when the wheel is moving such pressure distribution is slightly altered due to the effect of the elastic hysteresis of the material which composes the rails and the wheels. In particular in this case the sum of elementary ground reactions is slightly shifted towards the movement direction of δ (Figure 12b).

2.5.1. Basics of rail vehicle motion

Motion characteristics of a rail vehicle are strongly influenced by active and passive forces under which the train is subjected to during its motion. Such forces can be listed as:

- *Active forces* F which can be represented by both the so called “*induced tractive effort*” F_{Ti} supplied by the locomotive or the power equipment of the multiple unit, and the braking force F_B . The former has the same direction of speed vector v , while the latter clearly is directed in the opposite direction.
- *Passive forces* R , better known as motion resistances which have an opposite direction with respect to the speed vector and contrast tractive effort during vehicle movements.
- *Inertial forces*.

If the vehicle is modelled as a mass point P , whose mass is the vehicle mass m it is possible to write the Newton’s formula for the vehicle:

$$F - R = m_e \cdot a; \quad (8)$$

Where m_e is the equivalent vehicle mass, namely the mass of the vehicle increased by a constant factor accounting for rotating masses, while a is the acceleration vector.

Anyway before going deep into the description of each one of the actions involved within rail vehicle motion, it is necessary to underline that due to the adhesion phenomenon the tractive effort between wheel rim and rail is upper bounded, and if such effort exceeds such limit the wheels will spin. In particular active tangential actions F , as well as the vertical load G are transmitted from the wheels to the rail by means of the aforementioned contact area A , thanks to the so called *adhesion phenomenon*. As seen before, the load G is distributed on A with pressure p , therefore it

is possible to write G as: $G = \int_A p \cdot dA$. In the same way if $t = dF / dA$ is the elementary

tangential action, acting on area A , the resultant of tangential forces is: $F = \int_A t \cdot dA$. A

wheel subjected to the vertical load G and the tangential action F will roll on the rail only if adhesion conditions are verified. In fact if F is increased until it overcomes the so called adhesion limit F_{ad} the wheel will stop rolling and will start spinning. Defining the adhesion coefficient as $\mu = F_{ad} / G$, the adhesion conditions are verified only if $F \leq \mu \cdot G$, since F_{ad} represent the maximum value of tangential forces that can be transmitted from the wheel to the rail. Obviously such condition must be verified by both the tractive effort and the braking force, in fact if this latter overcomes the adhesion limit, the wheel will tend to be blocked while rubbing against the rail, also increasing the fatigue rate of the wheel rim.

However many studies have been conducted on measurements and determinations of the adhesion coefficient within railway systems in different conditions. In particular as can be easily understood, the value of the adhesion coefficient strongly depends on the vehicle speed as well as on the conditions of the contact area between wheels and rails, e.g. if the surfaces are wet (due to rain, snow, etc.) or dry. For example *Muller* (1953) identified the following law for calculating the adhesion coefficient:

$$\mu = \frac{\mu_0}{1 + 0.011 \cdot v} \quad (v \text{ in km/h}), \quad (9)$$

where $\mu_0 = 0.25$ for wet rails and $\mu_0 = 0.33$ for dry rails.

Curtius and Kniffler (1943) instead found the experimental results illustrated in Figure 13, where the continuous line is interpreted by:

$$\mu = \frac{7.5}{v + 44} + 0.161 \quad (v \text{ in km/h}). \quad (10)$$

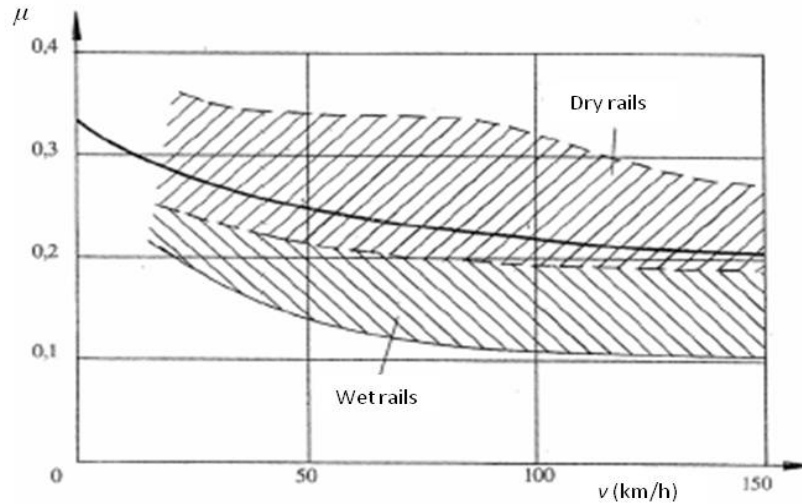


Figure 13. Experimental results from Curtius and Kniffler on measurement of rail adhesion coefficient μ with respect to speed values and weather conditions

2.5.2. Active Forces

As said before active forces are represented by both the tractive effort during acceleration phases, as well as the braking force within deceleration phases. In particular the “induced tractive effort” F_{Ti} is generated by the engine torque and transferred tangentially to rails through the contact area. Moreover such effort will depend on the vehicle speed and the curve which describes the tractive effort at wheel rim versus vehicle speed, constitute the mechanical characteristic of the traction unit.

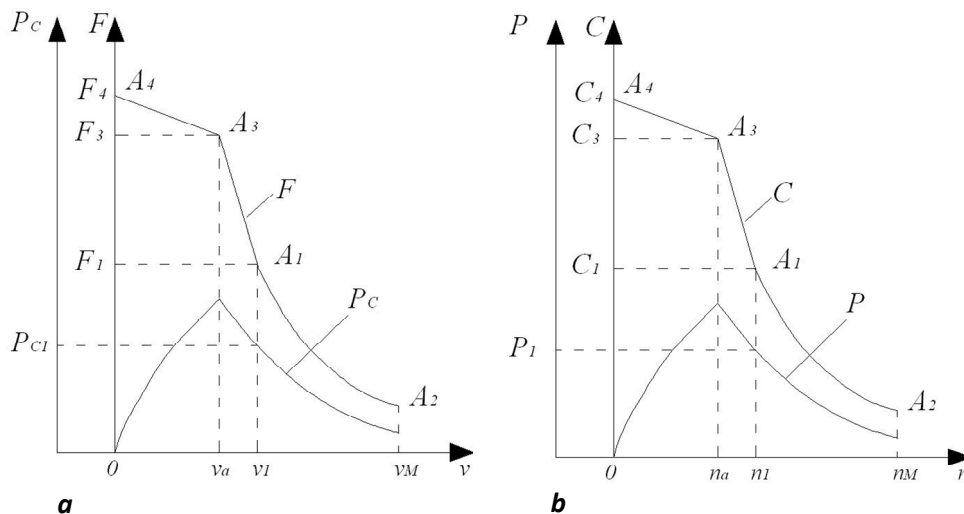


Figure 14. Characteristic curve of Tractive Effort F (a) , Torque C (b) and relative Power consumption of a DC Traction Unit.

In Figure 14a the characteristic curve of a traditional electric DC traction unit is depicted.

The power generated by the unit at the wheel rim can be calculated as:

$$P = F \cdot v; \quad (11)$$

However, three points of the characteristic curve must be commented: point $A_1(v_1, F_1)$ represents the nominal working of electrical engines, to the point A_2 corresponds instead the maximum speed v_M that electrical engines can reach, and point $A_3(v_a, F_3)$ corresponds instead to the “usage limit” of the characteristic $F(v)$. Within the so called “start range” delimited by speed values ranging from $0-v_a$, the tractive effort can reach values ranged between F_3 and F_4 . Indeed, the latter value of tractive effort and therefore the link $A_3 - A_4$ of the characteristic curve, can be reached only by means of an opportune regulation of the electrical engine during initial activation phases and compatibly with adhesion limits and the maximum engine torque. For these reasons typical characteristic curves of electrical engine assume that for $v < v_a$ the tractive effort at the wheel rim is constant and equal to F_3 .

Considering a rail vehicle equipped with N identical traction engines having the mechanical characteristic curve torque/rotation speed represented in Figure 14b, it is possible to define the *transmission ratio* ρ as : $\rho = \omega / \omega_r = n / n_r$, where ω and ω_r are respectively the angular velocities of the drive shaft and of the wheels with diameter D , while n and n_r represent the corresponding rotation speeds (in s^{-1}). Obviously it is always verified the condition $n > n_r$ and therefore $\rho > 1$. Given such relationships the translation speed of vehicle wheels (and therefore of the vehicle itself) can be expressed as:

$$v = \omega_r \cdot D / 2 = \frac{\omega \cdot D}{2\rho} = \frac{\pi \cdot n \cdot D}{\rho} = \frac{n}{k_n}, \quad \text{where } k_n = \frac{\rho}{\pi \cdot D}; \quad (12)$$

Moreover the relation between the torque applied to the wheels C_r and the torque applied to the drive shaft C can be written as: $C_r = \eta \cdot \rho \cdot C$, where η represents an efficiency factor taking into account for losses due to frictions between engine gears. As a consequence the tractive effort at wheel rim for N driving wheelsets can be calculated as:

$$F = N \cdot \frac{C_r}{D/2} = \eta \cdot 2 \cdot N \cdot \frac{\rho}{D} = k_f \cdot \eta \cdot C, \quad \text{where } k_f = 2 \cdot N \cdot \frac{\rho}{D}; \quad (13)$$

Furthermore, in order to adapt a locomotive to different operational tasks (e.g. passenger service: higher speed values, or freight trains: higher tractive efforts), different transmission rates are included within the same unit (Figure 15).

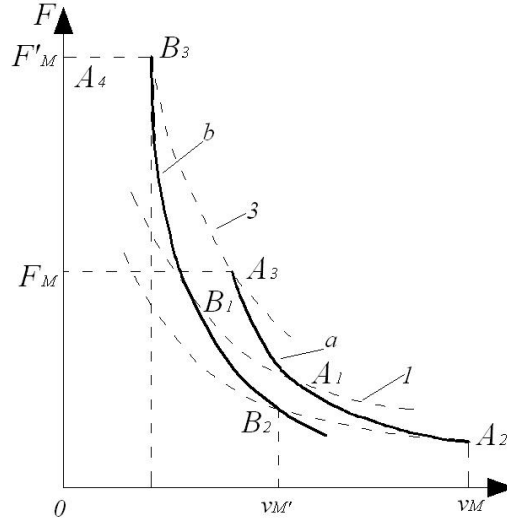


Figure 15. Characteristic Curve of a traction unit with different transmission rates: curve *a* for fast operations (passenger service) , curve *b* for heavy operations (freight trains)

In this way a single traction unit can be used for different roles according to the operator needs, which can change the transmission rate by means of apposite on-board items, of course movable only if the unit is stopped. The mechanical tractive effort-speed characteristic curve of the traction unit can be however analytically described by a group of hyperbolic or parabolic formulas, each of which is defined for a determined speed interval from v_k to v_{k+1} :

$$F_{Ti}(v) = c_{0,k} + c_{1,k} \cdot v + c_{2,k} \cdot v^2, \quad v_k < v \leq v_{k+1} \quad (13a)$$

$$F_{Ti}(v) = c_{h,k} / v, \quad v_k < v \leq v_{k+1}. \quad (13b)$$

For what concerns instead the braking forces, these can be applied directly to the wheel rim through different technologies that can be mainly branched in: *mechanical braking systems*, *electrical braking systems* and *electromagnetic braking systems*. Actually, any rail vehicle is equipped with a mechanical braking system whose main action is simply due to friction forces produced by pressing apposite brake blocks (usually made of cast-iron) on the wheel rim, or with gaskets positioned on opportune plates directly mounted

on the wheelsets. The braking command is usually based on compressed air systems equipped with an on-board pneumatic installation generally working at the average pressure of 7-9 bar, and supplied by apposite compressors directly located on the traction unit. However, mechanical braking systems must respond to the following characteristics:

- *Continuity*: namely the braking system must be controllable from a unique control place. To this purpose in fact a pneumatic conduit covers all the coaches composing the entire rail vehicle, in order to contemporary transfer to each one of them the braking command. In particular, for passenger rail vehicles where the braking readiness is a fundamental issue, an electrical-pneumatic braking system is used, where the braking action is always performed by a compressed-air system while the braking commands are transmitted electrically.
- *Automaticity*: in the sense that the braking actions must be applied immediately if the conduit is subjected to a failure due to a breaking of the conduit itself or of items which link together the several coaches of the vehicle. For this reason the braking action determines a decrease of the pressure within the conduit.
- *Scalability*: namely the braking action must be adjustable according to train movement necessities.
- *Inexhaustibility*: in the sense that the braking system must be always efficient also within long time periods. In fact, the loss of braking effort allowed for an efficient system does not overcome the threshold of 10-15%.

Since electrical engines are reversible in the sense that the engine rotor (and therefore the drive shaft) can be spun both in a way and in the opposite way, for DC traction units is possible to apply an electrical braking action directly on the wheelsets, which substitutes mechanical braking actions in particular when train speeds become higher than 160-180 km/h, therefore reducing strongly the fatigue of brake blocks and wheel rims. Anyway the safety of the braking action is always guaranteed by mechanical braking systems which in fact must always be applied during ordinary service when the effect of electrical braking disappears for low speed values (15-20 km/h), or when a failure to the electric braking system occurs.

Sometimes in addition to the mechanical and the electrical braking systems, rail vehicles may be equipped with a further electromagnetic braking system which is constituted by track brakes where the braking element is pressed by magnetic force to the rail, and therefore the braking action is performed by friction, and not the magnetic effect directly. However the braking effort supplied is independent on the adhesion and can be calculated as $F' = f' \cdot F_0$, where f' represents the friction factor and F_0 is the force developed by the magnetic attraction between rails and magnetic elements mounted on the wheelsets. When added to the other braking systems, electromagnetic brakes consent to reach very high values of deceleration rates (until 2 m/s^2). Since for passengers vehicles such deceleration values are too high to stand, especially for comfort and overall safety reasons, electromagnetic braking is usually activated only for emergency braking when sudden obstructions on the track or adverse adhesion conditions do occur.

For rail vehicles therefore a distinction must be done between ordinary service braking which is realized to meet comfort and safety requirements especially for passenger trains, and emergency stop which is instead automatically or manually (it depends on the reason for why it is activated) applied to keep safety conditions within dangerous circumstances. In both cases, after a brief initial transient phase where the braking action rapidly increase (with a jerk depending on the category of the considered train), the deceleration value can be considered as constant and varies according to the kind of service delivered by the train as well as by the circumstances within which it is activated. Normally the following typical values for braking deceleration a_f are given (*Brunger & Dahlhaus, 2008*):

- $a_f = 0.525 \text{ m/s}^2$, for suburban trains (service braking)
- $a_f = 0.375 \text{ m/s}^2$, for passenger trains (service braking)
- $a_f = 0.225 \text{ m/s}^2$, for freight trains (service braking)
- $a_f = 0.7 \text{ m/s}^2$, for suburban trains (sharp braking)
- $a_f = 0.5 \text{ m/s}^2$, for passenger trains (sharp braking)

- $a_f = 0.3 \text{ m/s}^2$, for freight trains (service braking)

In general, the deceleration value for service braking is about 0.75 times the value of the sharp braking. The braking effort therefore is calculated as : $F_B = m_e \cdot a_f$, where m_e is the equivalent mass of the vehicle.

2.5.3. Vehicle Motion Resistances

Passive forces, commonly known as “motion resistances” have an opposite direction with respect to the speed vector of the rail vehicle and they are due to several physical matters. In fact such resistances are generally decomposed in the sum of elementary resistances which can depend both upon vehicle and line characteristics. In particular the first kind of resistances mainly caused by the features of the vehicle itself can be distinguished amongst:

- Air resistance (with quadratic dependence on the train velocity relative to the wind)
- Rolling resistances caused by wheel rims, axle-boxes, adhesion, and similar reasons (parts of which are constant, and parts which have approximately linear dependency on the velocity)

In practice, these resistances are generally described by parabolas with coefficients r_i which depend on the train characteristics and the wind speed:

$$R_R(v) = r_0 + r_1 \cdot v + r_2 \cdot v^2; \quad (14)$$

As for a special train configuration these coefficient need to be appositely calculated from the data known about the train. There are heuristic formulae to estimate the resistances from the train characteristics as described in the following sections.

Traction unit resistances

For the traction unit (including multiple units), the resistance R_{TR} is usually described with given parameters a_0 , a_1 , and a_2 or a_{2r} for the following formula:

$$R_{TR}(v) = g \cdot m_T \cdot (a_0 + a_1 \cdot v) + a_2 \cdot v^2 + a_{2r} \cdot v_r^2, \quad [\text{N}] \quad (15)$$

Where m_T [Kg] is the mass of the traction unit, v [m/s] represents the speed of the vehicle, while v_r [m/s] is the relative speed between air and the rail vehicle usually assumed as $v_r = v + 4.17$ m/s (i.e. headwind of 15 km/h).

Other similar formulae are used in different countries, and in particular in Italy the empirical formula given by the Italian railway operator FS for computing resistances of the traction unit due to air viscosity is the following:

$$R_{TR}(v) = 4.2 \cdot G_T + 0.72 \cdot v^2, \quad [\text{kN}] \quad (16)$$

Where G_T is the total weight of the traction unit [kN], while v is the train speed [km/h].

Vehicle resistances for Passenger Trains

For passenger trains, the parameters of equation (15) are mainly described by: the mass of the vehicle m_W [kg], a factor c_b for the number of axles which can be assumed as 0.0025 for vehicles with 4 axles, 0.004 for those with 3 axles and 0.007 for those with 2 axles, the number of coaches n_W , and a value A_f [m²] which stands for the cross-sectional area of the vehicles weighted with their aerodynamic behaviour. Normally this value is assumed as 1.45 m².

Using the formula of *Southoff*, the relationship between the vehicle speed v [km/h] and the passenger vehicle resistance R_W [N] has theoretically and experimentally been determined as follows (*Southoff, 1932*):

$$R_W(v) = (1.9 + c_b \cdot v) \cdot \frac{g \cdot m_W}{1000} + 4.7 \cdot (n_W + 2.7) \cdot A_f \cdot \left(\frac{v + 15}{10} \right)^2 \quad (17)$$

Vehicle resistances for Freight Trains

For freight trains the formula of Strahl (Strahl, 1913) can be used to approximate the vehicle resistance R_{FW} [N] dependent on the vehicle speed v [m/s]:

$$R_{FW}(v) = 1000 \cdot m_W \cdot g \cdot (c_a + (0.007 + c_m) \cdot (3.6v)^2 / 100) \quad (18)$$

Where m_W [kg] is the mass of the wagons, c_a is a coefficient for axle adhesion (1.4 for roller bearings and 2.0 for older plain-bearing axle-boxes), and a value for air resistance c_m dependent on the kind of wagons (in particular it can be assumed to: 0.05 for mixed trains, 0.032 for full train loads, 0.04 for closed wagons and 0.1 for empty open wagons).

Line Resistances

As regards resistances caused by track characteristics (the so-called line resistances) it is necessary to specify that such passive forces are mainly due to line gradient as well as the curvature radii of the track as previously introduced within the preceding sections.

In particular for what concerns resistances due to line gradient, it is possible to say that the weight force G [N] on a slope of angle α can be described by:

$$G = m \cdot g \cdot \sin \alpha .$$

However since the gradients of railways are very slight, it is possible approximate $\sin \alpha$ to $\tan \alpha$, which is usually measured as n in per thousand [‰]. Measuring the mass m of the complete train (traction unit + coaches) in kg, the gradient line resistance R_g [N] comes to:

$$R_g = 1000 \cdot g \cdot m \cdot n , \quad (19)$$

For what concerns resistances R_c [N] due to sharp curves instead it is possible to assume the following formula:

$$R_c = 1000 \cdot g \cdot m \cdot \frac{700}{r} , \quad (20)$$

Where r represents the radius of the track curve expressed in [m]. It is immediate to understand that such resistance can be left out from the calculation on (nearly) straight lines, and that a curve with a radius of 700 m, generates the same resistance as a slope of 1‰, so the error cause by leaving out R_c is relatively small.

Moreover, the influence of the air resistance as a function of the cross-section and speed in tunnels should be regarded. However, there is no formula of general acceptance but the main significant effect is caused by trains meeting each other in a tunnel, but information on this is often not available when estimating the running times.

2.5.4. Vehicle Motion Formula and Rotating masses

As seen from the previous section it is possible to group all the vehicle resistances in a parabola formula dependent on the velocity as expressed in (15), with some easy binomial transformation, if the characteristics of train and line are known. The difference between tractive effort at wheel rim $F_T(v)$ and the sum of all the

aforementioned resistances $R = R_{TR}(v) + R_W(v) + R_g + R_c$, is left for accelerating the train.

However the train contains some rotating parts which consume some of the effort. The usual way to take care of this effect is a mass factor f_ρ for each part of the train. For the traction unit, the factor $f_{\rho T}$ should be given with the engine data (but usually it is not so distant from 1.09); for passenger vehicles and freight wagons instead it is possible to assume $f_{\rho W} \approx 1.06$. For the complete train therefore it comes out to:

$$f_\rho = \frac{(f_{\rho T} \cdot m_T + f_{\rho W} \cdot m_W)}{(m_T + m_W)}, \quad (21)$$

where m_T and m_W are respectively the mass of the traction unit and the mass of the wagons of the train.

At this point it is possible to specify the Newton's motion formula for the train as :

$$F_{Ti}(v) - (R_{TR}(v) + R_W(v) + R_g + R_c) = f_\rho \cdot m \cdot \frac{dv}{dt}, \quad (22)$$

The estimation of train running times can be therefore realized by using equation (22) and mainly following these tasks:

- Interpreting all given data of train and infrastructure
- Partitioning the route into pieces of equal characteristic and behaviour
- Solving the differential equation (22) for each homogeneous route partition.

2.6. Signalling equipments and other computer-based systems

As said before, signalling system represents a fundamental element within railway networks, since it is the responsible for the regulation of train movements respecting high-safety criteria, both for a single train run and for its interactions with other trains. Signalling equipments in fact assures safe train movements within the open track regulating the occupation of block sections amongst consecutive trains, or establishing the overcoming rules for opposite and/or intersecting train routes within network joints areas. But also within station areas, the so-called interlocking systems as well as other computer based movable network elements (points, switches) guarantee safe train station movements contemporarily assuring trains follow their own route and reach the

established platforms for fulfilling their operations (boarding/alighting of passengers, loading/unloading of freights, waiting for train connections, etc.).

Since the human being is the weakest element in railway safety it is necessary to equip both rail vehicles and the track with opportune automatic systems able to increase safety levels protecting against driver's errors or supervising his actions. Such systems are usually divided in Automatic Train Protection (ATP) and Automatic Train Operation (ATO) systems. The first kind of equipments is mainly related with keeping safety conditions during operations, and intervene automatically in case of driver's errors, such as the violation of a restricted signal, or of a restricted speed limit, preventing therefore events which would compromise the overall system safety. The ATO systems instead supervise or help the driver in carrying out more efficiently some actions related with train operations. They are more linked to efficiency of operations rather than to safety, and numerous examples are available throughout the world such as ATO for correctly docking the train towards station platforms or for automatic train driving (within partially automated or driverless systems).

2.6.1. Automatic Train Protection

As said before such systems have the main task of assuring protection against driver's error intervening automatically or semi-automatically in order to avoid dangerous conditions, guaranteeing the safety of the overall system and its customers. Modern technologies, have permitted the introduction of the so-called "cab-signalling" which integrates line side signals with in-cab indications to the driver or in some cases dispense with trackside signals altogether. Since cab-signalling constitutes now a fundamental part of automatic train protection systems, it is worth describing its main functions and then going deep in the heart of the different kind of ATP items.

The main functions of cab-signalling are:

Non-selective warning signals (mainly audible): whenever the train passes a certain position, e.g. the location of a distant signal, a warning tone sounds to direct the driver's attention to the trackside signals, independently from the signal aspect. No information connection between the trackside signal and the train protection system needs to be provided for this function which is applied in old train protection systems.

- **Selective warning signals** (again mainly audible): the audible signal is applied selectively in cases which imply restrictions for the driver, such as a restricted signal aspect which require the start of a braking action.
- **Visual repetition of trackside signals**: the aspect of trackside signals in advance is repeated in the cab during the train's passage within a certain block section, or while the train is within a defined partial section in the surroundings of the trackside signal. This type of cab-signalling can replace line side signals but often it is used additionally, since cab signals are visible to the driver in any kind of weather's condition. Anyway the cab signal does not provide any more information than the trackside signal, therefore the driver is responsible for estimating the braking requirement.
- **Continuous static speed information**: This kind of cab-signalling repeats the signal aspects within the cab and displays also the permitted speed under consideration of all restrictions (Figure 16). In addition, speed restriction warning information can also be displayed, but the driver is still responsible for estimating the braking curve. In several modern systems this type of cab-signalling replaces trackside signals.
- **Dynamic speed information**: Based on the static speed information, braking patterns are calculated on the train and/or in the trackside equipment. The technical system displays a guidance continuously which must be not exceeded momentarily in order to comply with the next target speed, to the driver (Figure 16). Information about the next braking target has to be present. This information can be either transmitted individually for each track section, or standardised by the uniform length of the section.

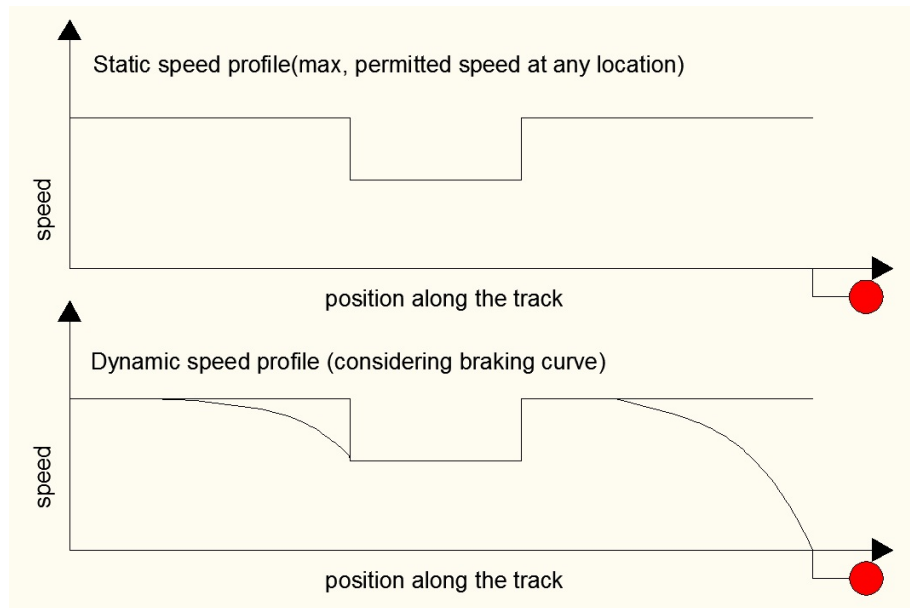


Figure 16. Functions of ATP systems: Static speed profile control or Dynamic speed control.

However the principal functions of ATP systems can be easily grouped within the following actions:

- **Check on driver ability.** Independently from the trackside information, the driver has to use an alertness device to guard against falling asleep or similarly, the so-called “dead-man’s handle”. Therefore it monitors the presence or the abilities of the driver during service.
- **Check on driver attentiveness.** In certain situations, such as passing a restricted signal, the driver has to acknowledge his attentiveness, by pushing for instance a special button. In this way in fact the danger from a driver failing to perceive a signal can be reduced significantly.
- **Train stop function.** The passing of a red signal is detected, which results in an immediate emergency stop. A particular issue is permissive driving, driving on written instruction or on an auxiliary signal. This is enabled either by additional override handles in the driver’s cab, by generally permitting the passage of the signal at very low speed or by a combination of these two methods.

Moreover, modern ATP systems provide a braking supervision which is applied when the train has to brake for a signal at danger or to comply with a speed restriction. Such systems supervise the braking process continuously or at certain points, in fact different

kind of ATP systems can be distinguished according to the type of braking curve supervision (Figure 17) in:

- *Supervision curve for the individual train*: which calculates and supervise for a single train and for the specific track layout, the braking curve which must be followed by the train to stop in rear of a stop signal. Advanced systems with digital data transmission like ETCS level 1 mainly use this method.
- *Brake supervision by standardized fragments*, where a stock of standardized fragments of brake patterns, differentiated by the speed level, proximity to the stop signal and/or train category, is provided by the system. Initiated by a trackside transmitter the proper fragment is selected. This is the typical solution adopted for spot transmission systems with low data volume such as the German system Indusi/PZB 90.
- *Staircase supervision*, where the supervision function has the shape of a staircase of speeds. This is the typical solution for systems with continuous transmission of signal aspects by coded track circuits where the same input data are valid for the whole length of track circuit.
- *Spot supervision*, where the speed is checked in form of multiple spots. The supervision speed obviously decreases from one checkpoint to the next in approach to a stop signal.

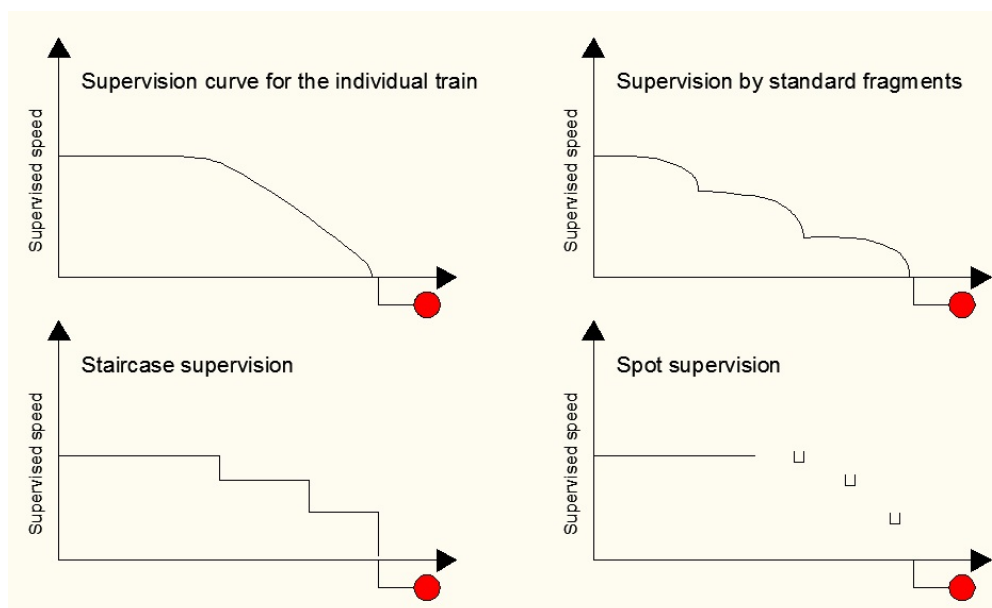


Figure 17. Different braking curve supervision of ATP systems

It is clear that each one of the described braking curve supervision system can be implemented only if the corresponding technical criteria of data transmission between the rail vehicle and the trackside is installed. For example the staircase supervision can be present only if there is a continuous data transmission like coded track circuit, while the spot supervision system is implanted when an intermittent transmission system is installed like for instance beacons or other kind of transponders. Generally, ATP systems can be in fact classified according to the kind of data transmission between the train and the trackside and the kind of function/information provided. In particular such systems can be distinguished in 6 main groups (Figure 18).

functions transmission	Attentiveness check, trainstop function and other without brake supervision	With brake supervision in different form, but without dynamic speed profile	Dynamicspeed profile
Intermittent	Group 1	Group 2	Group 4
Continuous	(3a) Group 3 (3b)		Group 5

Figure 18. Classification of ATP systems with respect to available functions and type of transmission

Group 1: Systems with intermittent transmission and without braking supervision

This systems do not have a braking curve supervision but instead have two other supervision functions: *a)* driver's attentiveness check at the signals showing caution (distant and combined signals), and/or *b)* trainstop functions. The gain in safety resulting from the application of these train protection systems is limited. This insufficiency has been in fact already demonstrated in the past for systems like the historic Japanese ATS-S, which only provided a check of attentiveness at signals showing "caution" independently from the current signal aspect. This system in fact reduced the number of accidents due to stop signal violation only by half, while 98% of the remaining accidents occurred after correct acknowledgment action by the driver (Kondo, 1980). Even when it can be assumed that selective acknowledgement check and the additional trainstop function increase the safety, in most cases these systems are not sufficient for modern safety requirements. The trainstop function without brake

supervision requires in fact overlaps long as the train braking distance, i.e. an empty block section must be always provided between two consecutive trains. However examples of such systems are: the mechanical Trainstop, the French Crocodile, the British AWS, and the Swiss Signum (see *Theeg, Vlasenko, 2009*, for references).

Group 2: Systems with intermittent transmission at low data volume and with braking supervision

For these systems, in addition to the attentiveness check and trainstop functions, the braking process is supervised in different forms, but without calculating a dynamic speed profile. For data transmission, resonant circuits are used in most cases. Each trackside resonant circuit is adjusted to a certain frequency out of a stock of defined frequencies, with the frequency coding information. These trackside resonant circuits can be moreover switched effective or ineffective, or they can be switched between different active statuses (different frequencies), depending on signal aspect. A disadvantage of many of these systems is that the ineffective (permitting) status cannot be distinguished from the absence of a trackside transmitter, which results in non-fail-safe behaviour of the system. For this reason, such systems are not suitable for cab-signalling and have to work in background as long as the driver is operating the train correctly. The driver must not be misled to rely on these systems. Example of such systems are the German Indusi/PZB 90 and the Japanese ATS-P (*Theeg, Vlasenko, 2009*).

Group 3: Systems with continuous transmission of signal aspects by coded track circuits

This kind of systems transmits the aspect of the trackside signals ahead to the train by means of coded currents running along the rails. The required track circuits are mostly also used for track clear detection and transmission of block information. The signal aspect ahead is repeated in the cab, often in simplified form. The supervision functions reach from simple acknowledgment checks up to braking supervision with standardised fragments. Basic advantages of systems of this group are the following:

- a) In contrast to most systems of group 1 and 2, these systems can be designed fail safe, so malfunction of the equipment leads to more restrictive indication in the cab.

- b) The train continuously receives the newest information in each position of the way. This prevents the driver from forgetting signal aspects and enables an immediate reaction of the system if signal aspect change.

Anyway, an important safety-reducing disadvantage is that, unless the length of the track circuits is standardised or additional transmitters for length information are provided, calculation of an adjusted braking curve is not possible. For such reason this type of system is integrated with intermittent transmission systems with the dynamic calculation of the braking curve (e.g. Russian SAUT system or ETCS level 1).

However examples of systems belonging to this group are: ALSN from Soviet Union, the Japanese ATC for High-Speed Rail and the Italian BACC, which will be later described in detail.

Group 4: Systems with intermittent transmission at high data volume and dynamic speed supervision

Due to fail-safe behaviour and the possibility to supervise the complete dynamic speed profile, these systems are a safe solution up to high speeds if the efforts for continuous transmission are not considered as necessary. Recently many of this kind of ATP systems have been developed and although their similar functional principles to a large extent, they are incompatible with each other due to different data coding and the amount of detail information and antennas. The main trackside transmission media are:

- a) Transponder balises, which work without track side power supply by using energy sent from the vehicle unit to send data telegrams back to the vehicle.
- b) Inductive loops with limited extension, which are usually powered from the trackside
- c) Locally limited radio transmission devices.

According to the data contents, transmission media can be divided into:

- Static data transmission media, whose information content is relative only to static characteristics of the track layout (track lengths, curvature radii, gradients), without giving any information on the aspect of trackside signals.
- Switchable transmission data media, which instead returns dynamic information about the current status of trackside signals aspect.

However the majority of ATP systems belonging to this group have trackside balises which both store static line data and dynamically communicate to the train the aspect of line-side signals. Examples of this kind of systems are: Ebicab (Scandinavia, Portugal, Brazil), ATB-NG (Netherlands), KVB (France) and the ETCS level 1 (International) which will be deeply described in the following.

Group 5: Systems with continuous transmission at high data volume and dynamic speed supervision

The basic difference between systems of this category and those belonging to group 4 is the continuous or quasi-continuous data link between track and train. Among systems of this category, the following technical transmission are applied:

- Coded track circuits, like Digital ATC (*Watanabe et al.* 1999) in Japan and TVM 430 (*Guilloux* 1990) which is applied on French and Belgian high-speed lines.
- Cable loops, as applied in the German LZB, mainly on high-speed lines.
- Radio transmission, implemented instead within ETCS level 2/3.

In contrast to most systems with intermittent transmission, information flow is centralised in most cases using a line-side control centre. The line-side control centre and the train computer are in most cases vital redundant microprocessor systems. The functions of the components can differ in detail between the systems. An important criterion to distinguish the systems is whether they are used as the only signal system on the respective lines or if they are used mixed with trackside signals. In the former case, system inherent fallback levels are provided, or driving on sight is the only fallback level for degraded mode operation. In the latter case, the cab signal system often enables shorter block sections and therefore higher line capacity than trackside signals. The assignment of functions to the interlocking system or the train control system is basically defined as follows on lines for mixed traffic:

- The interlocking functions including track clear detection, which are needed for all movements on the line, are assigned to the interlocking system.

- The particular cab signalling and train protection functions, which are only applicable to equipped trains, are assigned to the train control system. In some cases, additional auxiliary functions for interlocking can also be carried out in the train control system, such as detecting the halt of the train for route release.

Resulting from this assignment of functions, route information has to be transmitted from the interlocking system to the trackside control centre. For functions such as sending the information about the halt of a train to the interlocking system, a bidirectional data connection is necessary, otherwise an unidirectional connection suffices. Systems like the French TVM as well as the ETCS level 2 that will be better depicted successively, belong to this group.

A unified European Train Control Systems: the ETCS levels

In Europe a large variety of different signalling systems has been observed during the years and throughout the countries. However, these differences in system features which often were very strong when passing from a country to another one, have prevented an international interoperability between countries of the European continent. In fact, a train which needed to operate between two different countries (for example linking Italian and Swiss cities) had to be equipped with both Italian and Swiss signalling items which consented the communication with the two different systems. Moreover, it was not so infrequent that train driver had to be substituted at the frontier (e.g. the Italian driver was substituted with a Swiss driver) since the differences in signal lights configuration and other line-side messages could be not clear for a driver from a different country, implying higher risks for the run. To eliminate all these inconvenient, European Community decided to create a standard signalling system which could have been usable throughout all country members, allowing therefore a complete interoperability amongst European countries. These form of standardization, would have permitted strong economic savings (preventing the installation of multiple signalling equipments on trains, as well as the change of train driver beyond the state line, etc.) but above all an efficient interoperable network which would have eased the communication between European countries.

This unification process started in 1990s thanks to the European Commission who initiated the process, UIC (International Union of railways) who defined the functional requirements of the system and Unisig (Consortium of the 7 largest European signalling

manufacturers) who specified the detailed technical solutions for the system. ETCS is currently being introduced on several railways in Europe. Problems in the introduction process are however the high investment value in the existing national system and the migration from the old national systems to ETCS requiring double equipment of lines and/or vehicles for longer time. Besides Europe, several countries outside Europe use ETCS on some lines. These are currently Taiwan, South Korea, PR China, Saudi Arabia, Turkey, India, Australia and Mexico (*Garstenauer & Appel 2007, Winter et al. 2009*).

In particular four different levels have been identified for ETCS systems: level 0, 1, 2, 3.

The term “**ETCS level 0**” describes the situation where a vehicle which is equipped with ETCS moves in an unequipped area. The supervision functions are limited to supervision of a constant speed, which is the minimum of the maximum train speed and a general nationally defined speed limit for level 0. For example in Italy, it is possible to identify an ETCS level 0 when a rail vehicle equipped with ETCS items, run on a track where a coded track circuit system such as the BACC is installed without any ETCS trackside equipment. In such cases in fact the driver must follow the signalling rules properly of the BACC, that it is not able to provide the calculation of an adjusted dynamic braking curve profile but only a staircase speed supervision. Therefore it needs to keep a signal overlap whose length tallies with the block section length, i.e. an empty block section must be present between two consecutive trains. However, as said before the BACC system is adopted within Italian conventional lines, but also an integrated application with an ETCS level 1 system has been implemented on several high-speed lines (in particular on the Rome-Florence line). Such ATP system belongs to Group 3 and it is therefore based on continuous transmission of signal aspects by means of coded track circuits (Figure 19). It is also equipped with a cab-signalling with three different aspect (red, yellow and green) plus an additional fourth aspect (red-yellow) adopted when speed restrictions due to the track layout must be applied (e.g. to cross over switches, network joints, etc.). The signals on the carrier frequency 50 Hz are frequency modulated with frequencies 4.5 Hz (270 minutes⁻¹; green signal ahead), 3 Hz (180 min⁻¹, yellow signal ahead: speed reduction), 2 Hz (120 min⁻¹, red-yellow signal and speed restriction due to diverging route ahead) and 1.25 Hz (75 min⁻¹; stop signal ahead). As track circuits have almost the same length (equi-block layout usually of 1350 m for

mixed traffic railways), it is possible to supervise the braking curve by means of a staircase speed supervision. A stop therefore will be announced to the driver two block sections before the red signal. However, since for high-speed lines the braking distance takes more than two track circuits, this system has been upgraded adding a further carrier frequency of 178 Hz which in combination with the codes based on 50 Hz gives nine speed steps for high-speed trains. This system is moreover downwards compatible in the sense that high-speed trains can run on conventional lines and conventional trains on high speed lines using only the 50 Hz code and at speeds not higher than 200 km/h.

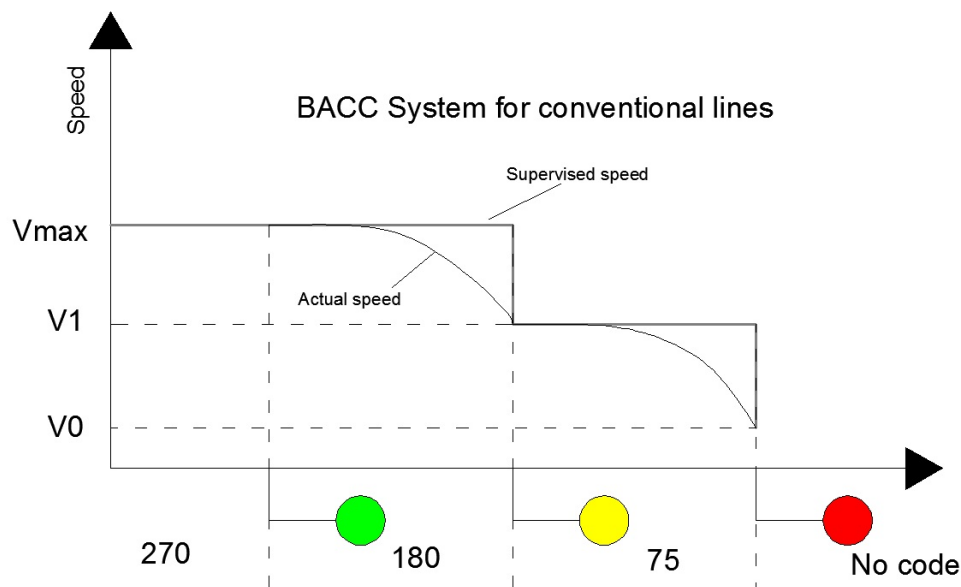


Figure 19. Working scheme of the Italian BACC signalling systems (based on coded track circuits)

ETCS level 1, instead belongs to group 4 since transmission of data between vehicles and trackside is intermittent and dynamic speed supervision is provided. The main transmission medium are transponder balises called “Eurobalise” which transmit, among others, movement authorities and profile data to the train (which is not individually known) when passing above the balise (Figure 20). In fact when the rail vehicle cross a balise this one is activated by means of electromagnetic induction principle and transmit data to the on-board European Vital Computer (EVC) which is the responsible for the calculation of the dynamic braking curve of the train as well as of its supervision. In particular the speed profile is controlled by the EVC which compares speed-position train data returned by on-board odometers with those calculated that must be followed to assure a safe braking procedure. Balises exchange with the EVC both fixed (track characteristics) and switchable data (dynamic signal aspects). In

particular to consent the transmission of switchable data a apposite electronic system is necessary on the trackside, the so-called LEU (Lineside Electronic Unit) which selects the data according to signal aspects. Balise area usually linked with each other, which means that most balise groups are announced by a previous balise group, enabling the detection of faulty balises by trainside distance measurement. Besides the balises, linear infill devices can be used locally to transmit changes of signal aspects beyond. These are Euroloop (cable loops in the rail) or radio infill units. The third type of information between track and train besides the normal and infill is repositioning information. This is used in cases when in some signal systems a trackside signal does not know the exact path the train is going to take to specify the information gaps after the branching. Based on the information received from the trackside and on the train data which include braking characteristics, the train computer calculates the dynamic speed limit which can be signalised to the driver in the cab signalling equipment and supervised. As level 1 provides continuous guidance functions by movement authority, trackside signals are optional, although used in most cases. Moreover Level 1 is used for conventional line or lines whose maximum speed is around 160 km/h.

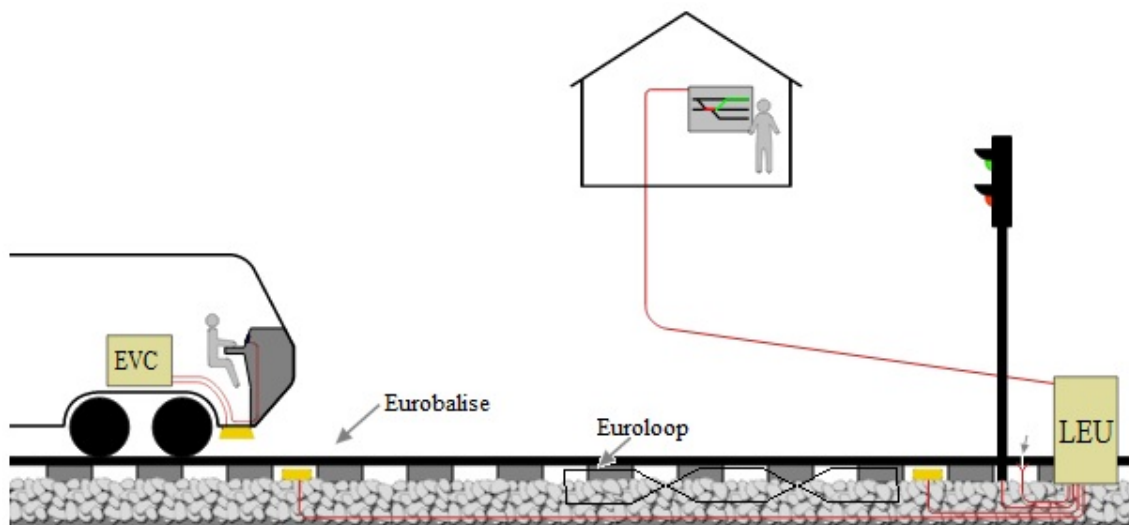


Figure 20. Working scheme of ETCS level 1 signalling system.

ETCS level 2 and 3 belong instead to group 5, since they provide a continuous data transmission between track and rail vehicles with a dynamic speed supervision. Information in fact are continuously and bi-directionally transmitted by Euroradio, a radio standard based on GSM-R. The central trackside unit is the Radio Block Centre (RBC). It is responsible for a longer section of line (in Italy it controls on average around sections of 70 km), stores the static data and obtains dynamic data like signal

and point positions from the interlocking stations in the area. In contrast to Level 1, trains are individually known in the RBC. The train requests new movement authorities in regular time interval (usually every 60 seconds) or at particular events. Here the balises transmit only static data inherent to physical track features (gradients, curvature radii, etc.) or reposition train odometers, while switching information relative to the current aspect of signals ahead and therefore movement authorities are directly communicated via radio. Within ETCS level 2 (Figure 21) systems line-side signals are generally removed since all the necessary information is already communicated through the radio system. Anyway train separation still need fixed block sections (highlighted by opportune block section boards) since the system lack of a continuous detector for train integrity.

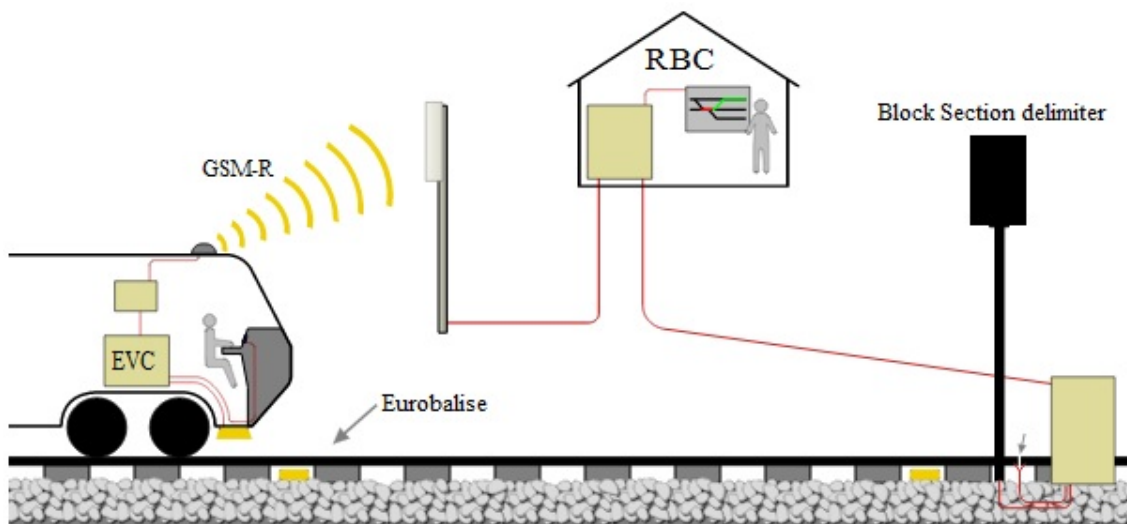


Figure 21. Working scheme of ETCS level 2 signalling system.

For what concerns **ETCS level 3** (Figure 22), it is necessary to specify that this system goes beyond the pure train protection functionality with the implementation of full radio-based train spacing. In fact it can be seen as an ETCS level 2 system, where train separation is no longer based on fixed-block sections but on the communication capacity of the radio system. This means that in this system fixed block sections are removed and with them all the track-release signalling devices (e.g. axle counters). In this way the so-called “moving block” principle would be applied, where train separation is based on the absolute braking distance spacing (as cars on the road). As in ETCS Level 2, trains find their position themselves by means of positioning beacons and via sensors (axle transducers, accelerometer and radar) and must also be capable of determining train integrity on-board to the very highest degree of reliability. However

Level 3 is currently under development since solutions for reliable train integrity supervision systems are necessary for its concrete implementation since fixed block section are completely removed. Furthermore another problem is due to the fact that such solutions are very complex to find as well as hardly suitable for transfer to older models of freight rolling stock.

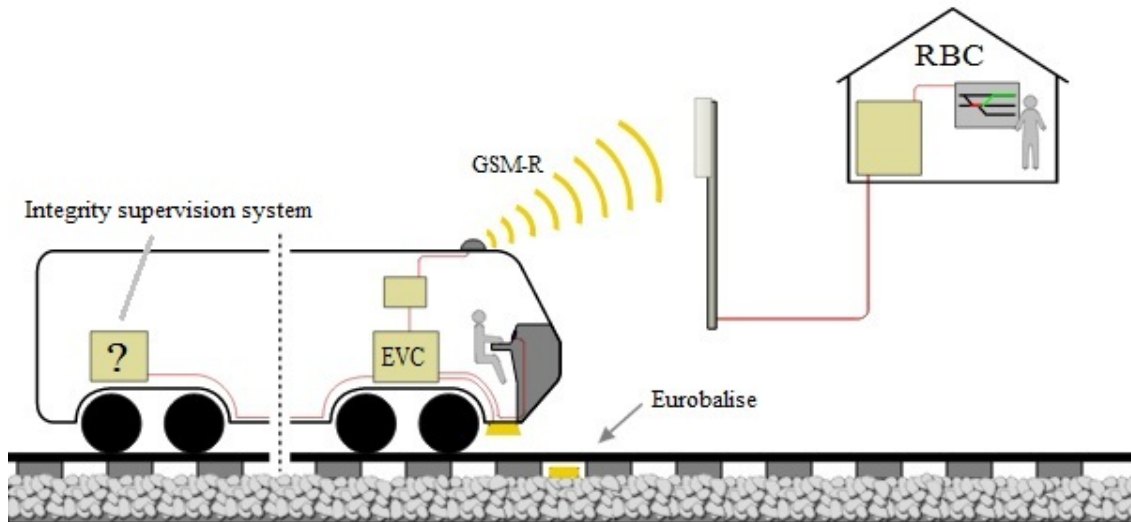


Figure 22. Working scheme of ETCS level 3 signalling system.

2.6.2. Automatic Train Operation

The ATO systems can be defined as items which support train drivers to better perform some operations (e.g. the correct positioning of the train towards station platforms when stopping) or to supervise train runs allowing for example the respect of signal limits on the track. In particular with complete dynamic speed profile present on the train, in principle, train operation could be automated. However reasons which obstruct complete automation are mainly the lack of ability of such systems to react to unforeseen situations such as obstacles on the track. Therefore, a further necessity for full automation of train operation is the continuous detection of external objects or their exclusion by barriers which cannot be passed either intentionally or unintentionally. This is very expensive on extended networks, but is practicable in some cases of metropolitan railways due to the limited extent of the network and the high density of traffic which justifies the economic investment. Complete protection is never possible against very rare events, such as an object falling from a passing aircraft onto the railway. Altogether, the following steps of automation can be distinguished:

- Manual driving without automation: the driver is fully responsible for driving. This is the case without train protection systems.
- Manual driving with technical supervision: This is the case of a train protection system supervising the driver and enforcing safety in case of driver's error.
- Partially automatic operation: This is the case when some tasks of driving regulation are assigned to the driver and others to automatic systems. An example is ATC on Japanese high speed lines, where the driver is responsible for acceleration and platform stopping and the automatic system for safety related braking processes. Other examples are several modern systems with calculation of dynamic speed profile where the driver can select between manual and automatic driving.
- Automatic driving with human supervision: Here the train is normally driven automatically, but the driver watches the track and can take actions in case of danger or technical failure. Although this would be technically possible in many modern systems, it is rarely done for psychological reasons: a driver whose only task in normal operation is watching the processes would not be able to act properly in emergency situations due to lack of attentiveness and driving practice (Yamanouchi, 1979). This can be overcome by giving the driver some positive tasks, as implemented on the Victoria Line of London Underground in 1968.
- Full automation: In these systems no driver who would watch continuously the track is present on the train. However, in some cases a person who is normally in charge of other tasks (such as selling tickets) can take control if necessary. Fully automatic driving is currently applied on some single metro lines (e.g. Copenhagen, Paris, Vancouver) and for special purposes such as airport shuttle trains.

Anyway down here some ATO systems dealing with the supervision of train stopping operations (ATO docking and starting) within station areas, as well as coupled ATO/ATP systems for the control of metropolitan train movements on the track, are described.

ATO Stopping, Docking and Starting

Watching the Figure 23, it is possible understand in an easy way the principles and modes of operation of a station ATO systems.

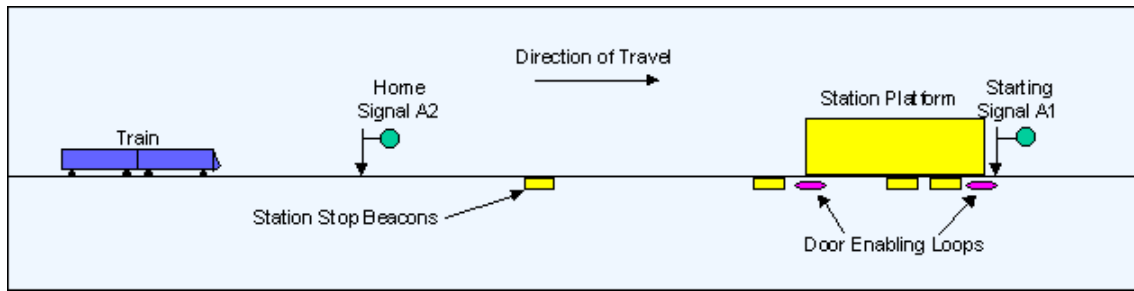


Figure 23. Scheme of a station ATO system for regulating operations of train docking at platforms in a Metro line (Railway Technical Web Pages, 2011).

The train approaches the station under clear signals so it can do a normal run in. When it reaches the first beacon - originally a looped cable, now usually a fixed transponder - a station brake command is received by the train. The on board computer calculates the braking curve to enable it to stop at the correct point and, as the train runs in towards the platform, the curve is updated a number of times (it varies from system to system) to ensure accuracy. London's Victoria Line, now 35 years old, has up to 13 "patches" checking the train speed as it brakes into a station. This high number of checks is needed because the on-board braking control gives only three fixed rates of deceleration. Even then, stopping accuracy is ± 2 meters. A detailed description of the Victoria Line's ATO system is here. Modern systems require less wayside checking because of the dynamic and more accurate on-board braking curve calculations. Now, modern installations can achieve ± 0.15 meters stopping accuracy - 14 times better.

In addition to providing an automatic station stop, ATO will allow "docking" for door operation and restarting from a station. If a "driver", more often called a "train operator" nowadays, is provided, he may be given the job of opening and closing the train doors at a station and restarting the train when all doors are proved closed. Some systems are designed to prevent doors being opened until the train is "docked" in the right place. Some systems even take door operation away from the operator and give it to the ATO system so additional equipment is provided as shown left. When the train has stopped, it verifies that its brakes are applied and checks that it has stopped within the door enabling loops. These loops verify the position of the train relative to the platform and which side the doors should open. Once all this is complete, the ATO will open the doors. After a set time, predetermined or varied by the control centre as required, the ATO will close the doors and automatically restart the train if the door closed proving circuit is complete. Some systems have platform screen doors as well. ATO will also provide a signal for these to open once it has completed the on-board

checking procedure. Although described here as an ATO function, door enabling at stations is often incorporated as part of the ATP equipment because it is regarded as a "vital" system and requires the same safety validation processes as ATP. Once door operation is completed, ATO will then accelerate the train to its cruising speed, allow it to coast to the next station brake command beacon and then brake into the next station, assuming no intervention by the ATP system.

ATO/ATP systems for Multi-Aspect Signalling in Metro networks

This system usually employed for metro systems to assure higher capacity values (therefore higher train frequencies) is described in Figure 24.

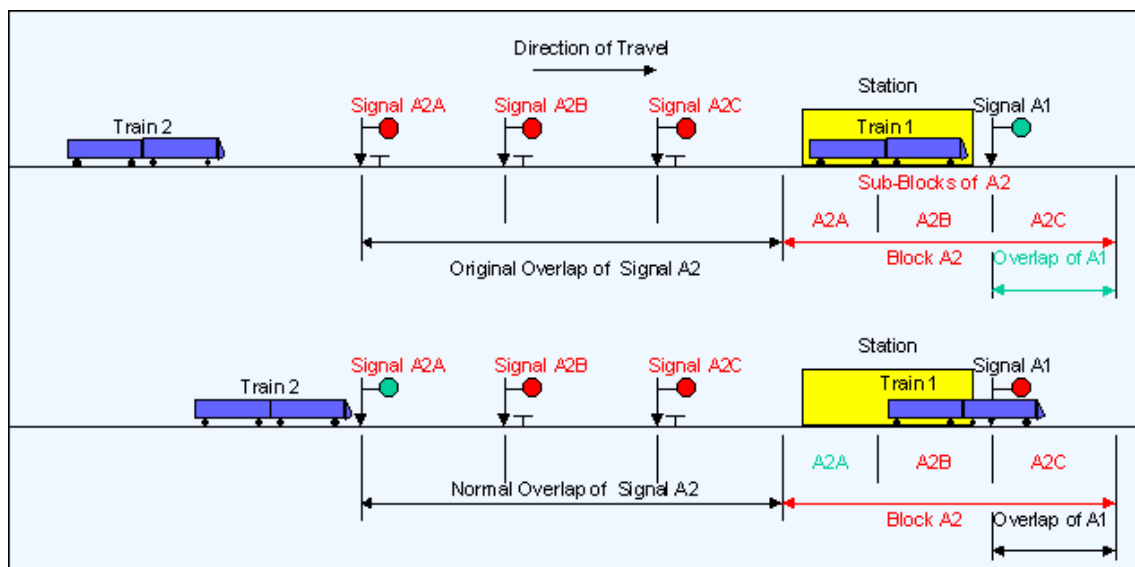


Figure 24. Scheme of a Multi-Aspect signalling system on a Metro line (Railway Technical Web Pages, 2011).

In particular where multi-home signalling is installed at a station, it involves the provision of more but shorter block sections, each with its own signal. The original home signal is Signal A2A and while Train 1 is in the platform, it will remain at danger. However, Block A2 is broken up into three smaller sub-blocks, A2A, A2B and A2C, each with its own signal. They will also be at danger while Train 1 is in the platform. Train 2 is approaching and beginning to brake so as to stop at Signal A2A. When Train 1 begins to leave the station, it will clear sub-block A2A first and signal A2A will then show green. Train 2 will have reduced speed somewhat but can now begin its run in towards the platform. Then after Train 1 has cleared two sub-blocks, A2A and A2B, two of the multi-home signals become consequently clear. In particular the starting signal is red as the train has entered the next block A1, while Train 2 is

running towards the station at a reduced speed but it has not had to stop. However, when Train 1 clears the overlap of signal A1, the whole of block A2 is clear and signal A2C clears to allow Train 2 an unobstructed run into the platform.

Therefore fixed block metro systems use multi-home signalling with ATO and ATP, where a series of sub-blocks are provided along the track and in particular within the platform area. These impose reduced speed braking curves on the incoming train and allow it to run towards the platform as the preceding train departs, whilst keeping a safe braking distance between them. Each curve represents a sub-block. Enforcement is carried out by the ATP system monitoring the train speed. The station stop beacons still give the train the data for the braking curve for the station stop but the train will recalculate the curve to compensate for the lower speed imposed by the ATP system (Figure 25).

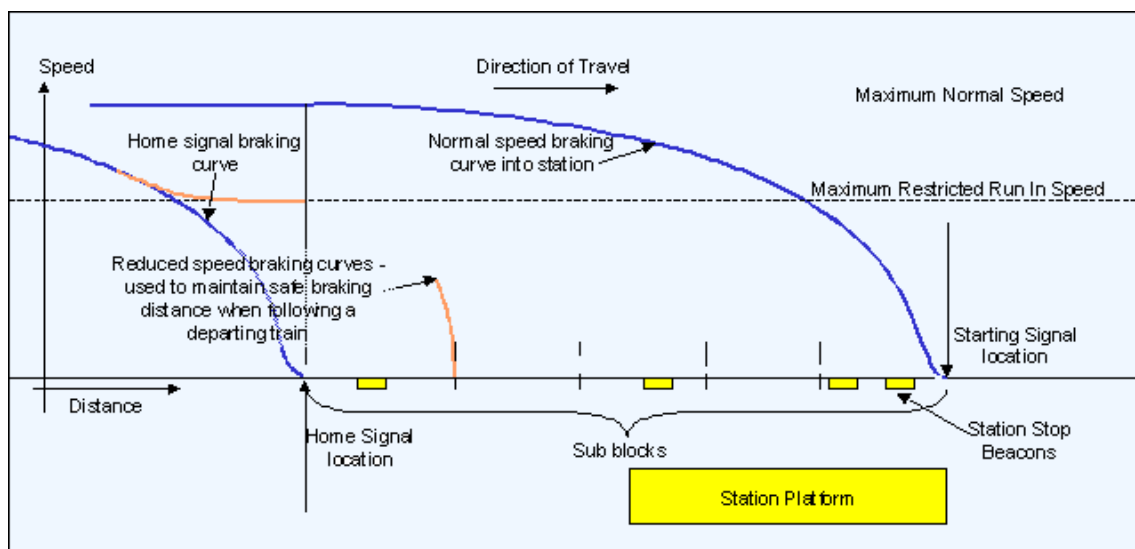


Figure 25. Supervised braking curve in a ETCS level 1 ATP system (Railway Technical Web Pages, 2011).

2.6.3. Interlocking Systems

Interlocking fulfils the function of “information processing”, in the sense that provide the central function which ensure that trains move safely in technical terms. To achieve this, the interlocking obtains information about track occupation (by rail vehicles and other objects) and the position of movable track elements. It then evaluates these information and permits movements via the signals. Amongst others, two basic principles are enforced technically by interlocking functions:

- A signal can only permit a train movement if all movable track elements are in proper position and locked (*dependence between points and signals*), and the elements must remain locked as long as they are being used by the train.
- With train spacing by fixed block, a train can only be permitted to enter a section which is clear of other rolling stock, and no other train may be permitted to enter that section.

Different kind of interlocking systems can be distinguished on the basis of their technology (mechanical, relay, electronic), although they have the same logic and apply the same principle of interlocking. Interlocking systems, are mostly installed within complex areas of the network, such as stations, complex joints, marshalling yards, where a large amount of different train paths must be managed at the same time. In such cases in fact, the probability of collision between trains, as well as the probability of derailments when crossing track discontinuity like switches (and points), becomes higher in these areas, therefore a simple unidirectional or bi-directional signalling framework is not more enough to safely control all these movements. To this purpose an opportune centralized system able to take into account and safely regulate all these processes is necessary, and that is why the interlocking system constitutes a fundamental component of railway signalling system.

Main interlocking functions are therefore addressed to guarantee:

- Protection against following movements,
- Protection against opposing movements,
- Protection against flank movements,
- Safety at movable track elements (switches).

In particular there are two basic principles regarding the methods of safeguarding the way, which can be distinguished and defined as follows:

Route. The whole path including the positions of movable track elements and track clear detection is only checked upon request, normally when setting the route before clearing the signal (and it is then supervised until the train enters the route). The principle “route” can provide almost all the four protective functions listed above, but it

also contributes to speed targeting at movable track elements. It can even incorporate level crossing and obstacle detection.

Block information. After a train has cleared the block section, the message confirming this is generated and transmitted to the entry point of the section. There, this information is stored, to permit the later entry of another train. The principle “block section” can only provide following and opposite protection. Therefore it is basically applicable to open line sections.

Historically this difference emerged from track clear detection being the observation of the signaller: within the control area of a signal box (an interlocking), the signallers who were responsible for setting the routes proved the track clear directly by sight before clearing the signal. In contrast, between two neighbouring signal boxes with a longer portion of line in between which was not visible from either of them, this principle could not be applied and other solutions had to be found. Here the clear status of the line section was concluded from information about trains entering and completely leaving the section.

However, the difference between the principles “route” and “block information” lost much of its operational importance with the introduction of continuous technical track clear detection. With this, all block system functions can also be provided by routes. Thus in recent years the principle of “block information” has been superseded in some countries, with the principle of “route” being used also in open lines. In most countries where train and shunting movements are determined separately, shunting movements are restricted to particular areas (e.g. station areas). The principle “block information” is basically applied to train movements. The principle “route”, in contrast can be applied to both train and shunting movements.

The logic which stays at the basis of interlocking functions, is the dependence between two or more infrastructural elements. Elements which can be interlocked are in fact constituted of:

- Movable track elements such as switches,
- Signals
- Other elements like level crossings

Moreover dependences amongst such elements can be distinguished according to different criteria, one of which is the number of elements to be locked: in fact the interlocking can be carried out by means of a dependence between only two elements, or by the dependence among three or more elements (the so-called “conditional locking”).

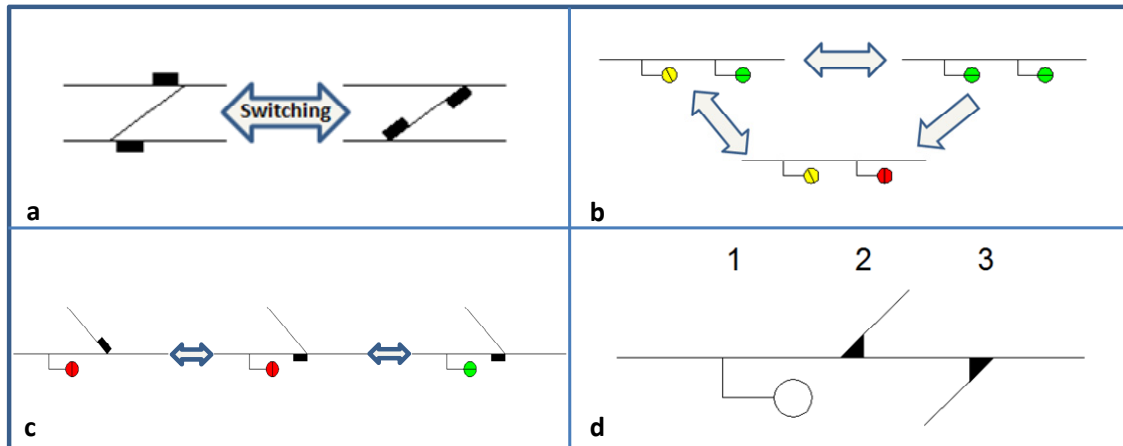


Figure 26. Types of dependencies amongst interlocking elements in a railway network: a) Coupling of two elements, b) Unidirectional locking, c) Simple bidirectional locking, d) Conditional locking.

According to the logical arrangements, element dependencies can be divided into the following elementary locking functions:

Coupling of two or more elements: where coupled elements are operated by the same operational element and can only be switched together in regular operation. The most typical case is that of two movable track elements giving flank protection to each other, such as two sets of points of a crossover or a set of points and a derailler protecting that set of points (Figure 26a).

Unidirectional locking of two or more elements: in this case it is possible to define both independent and dependent elements. The independent ones can be moved freely (unless locked by other functions). The dependent element can only be set to a certain position if the independent element is also in a certain position, and leaves this immediately when the independent element leaves its position. With more than two interlocked elements, the position of the dependent element depends on the combination of positions of the independent elements. For example the most typical case is the unidirectional locking which implies a dependence between main and distant signals (Figure 26b). In fact the distant signal can only show a proceed aspect if the main signal

is green. Therefore the main signal is the independent element since it changes its aspect according to the occupation state of the section that it protects, while the distant signal is the dependent part of the locking since its aspect is directly dependent on the main signal aspect.

Simple bidirectional locking: where two or more elements are interlocked that way so that one combination of positions is impossible and each element is locked if the others are in these respective positions. A typical example is constituted by the dependence between points and signals (Figure 26c). The signal is locked in the Stop position if the points are diverging, and the points are locked in the straight position if the signal shows a proceed aspect. This means that a certain combination of positions like signal at Proceed and points diverging is impossible to obtain for safety reasons.. The elements therefore have to be switched in a defined sequence: first, the points have to be straight, and then the signal can be cleared. Thus this kind of locking is also called sequential locking.

Conditional bidirectional locking: whose main logic is similar to the previous simple bidirectional locking, but here the number of involved elements is more than two. A simple example can be made looking at the Figure 26d, where signal 1 can only be cleared if there is a safe path ahead. One condition is that points 2 must be locked in either end position, which is in fact an OR-combination of two simple locks, one for each end position. However, if points 2 are straight, there is another safety condition: points 3 must also be locked straight. The commonly used term in Western European countries “conditional locking” (*Retiveau 1987, Such 1956*) means in the example of Figure 26d: “if points 3 are in the diverging position, then signal 1 at Proceed locks points 2 in the diverging position (and points 2 straight locks signal 1 at Stop)”. This is equivalent to “Signal 1 at Proceed, points 2 straight and points 3 diverging cannot occur at the same time”. Combinations of several simple and conditional locks form the basis for cascade route locking.

Anyway, AND/OR combinations of these four different elementary locking functions give rise to all complex interlocking functions and the formation of specific train routes and paths, required to manage dense railway traffic within station areas or complex network joints.

Route formation and all interlocking function can be carried out both locally, acting directly on local interlocking systems, and via remote control setting train paths and routes of interlocking areas from a distant control centre. For the first kind of control it is possible to talk about decentralized operation, while the second type deals with centralized operation. As said before within decentralized operation, train movements are controlled by local interlocking stations (Figure 27). The operators of neighbouring interlocking stations communicate to each other by means of telecommunication, mostly simple telephone connections. All communications between the local interlocking stations and all train movements are registered by the operators in train records. On European railways, movement authorities are issued by local operators. Because the local operator is the authority person, the dispatcher is only responsible for watching the traffic and for solving scheduling conflicts to avoid delays and congestion. Thus the dispatcher supports local operators in an efficient operation.

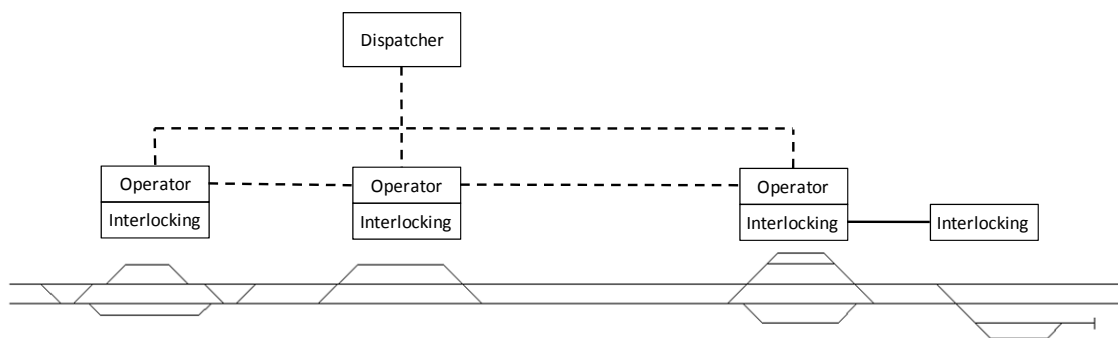


Figure 27. Decentralized railway operation management with local interlocking stations.

In Centralized Traffic Control (CTC) instead, all points and signals inside the controlled area are directly controlled by the dispatcher (Figure 28). All train movements are controlled by signal indications. The local interlockings are remote-controlled without local staff. In CTC territory, all main tracks must be equipped with track clear detection. CTC technology has a long tradition on railways that operate long lines in territories with a very low population density and long distances between stations.

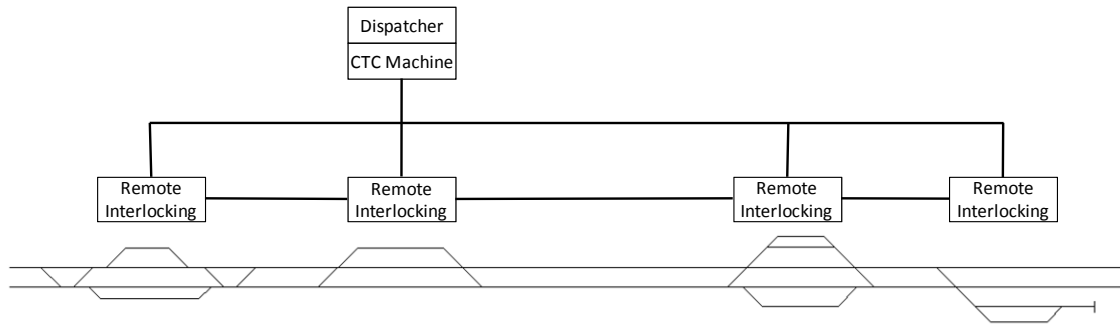


Figure 28. Centralized railway operations with remote controlled interlocking stations.

The described functions are provided by all interlocking systems independently from their specific technology. However a brief introduction to the different kind of interlocking technologies used throughout time and different countries is useful. In particular it is possible to distinguish, mechanical, electro-mechanical, relay (or electrical), and electronic interlocking systems.

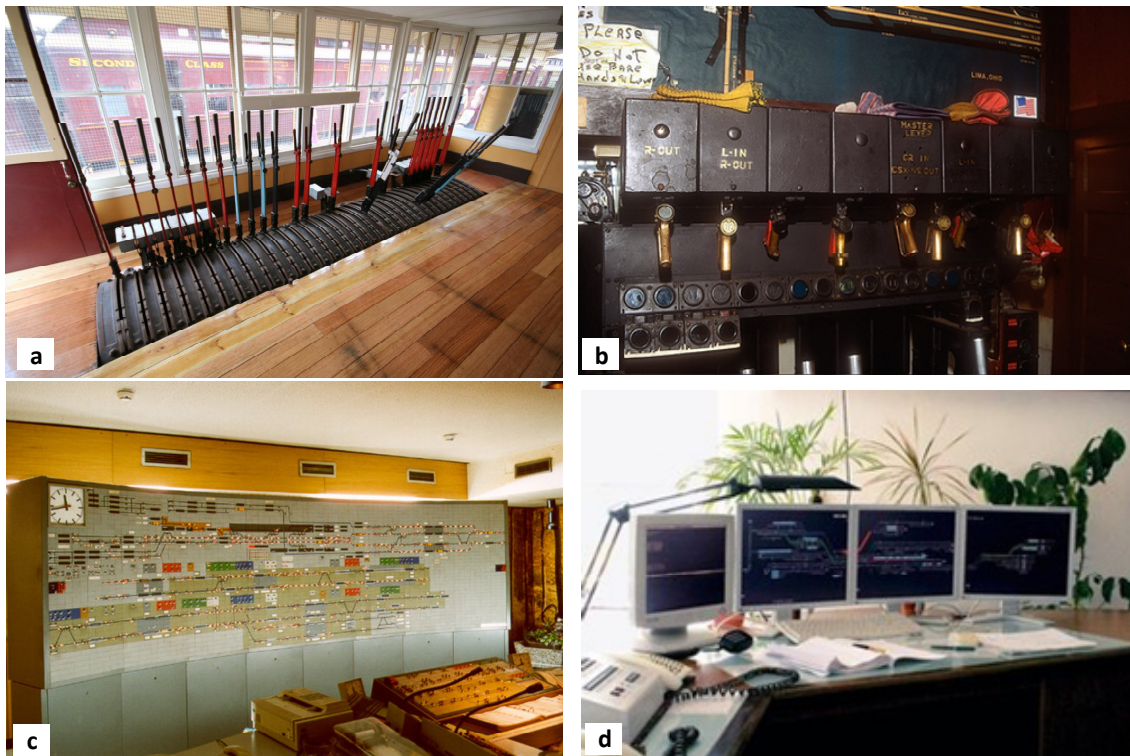


Figure 29. Different technologies of Interlocking systems: a) Mechanical, b) Electro-Mechanical, c) Relay, d) Electronic.

Mechanical interlocking

In mechanical interlocking plants, a locking bed is constructed, consisting of steel bars forming a grid (Figure 29a). The levers that operate switches, derails, signals or other appliances are connected to the bars running in one direction. The bars are constructed

so that, if the function controlled by a given lever conflicts with that controlled by another lever, mechanical interference is set up in the cross locking between the two bars, in turn preventing the conflicting lever movement from being made. In purely mechanical plants, the levers operate the field devices, such as signals, directly via a mechanical rodding or wire connection. The levers are about shoulder height since they must supply a mechanical advantage for the operator. Cross locking of levers was effected such that the extra leverage could not defeat the locking (preliminary latch lock). Generally this kind of interlocking can be only locally operated.

Electro-Mechanical interlocking

Power interlocking may also use mechanical locking to ensure the proper sequencing of levers, but the levers are considerably smaller as they themselves do not directly control the field devices (Figure 29b). If the lever is free to move based on the locking bed, contacts on the levers actuate the switches and signals which are operated electrically or electro-pneumatically. Before a control lever may be moved into a position which would release other levers, an indication must be received from the field element that it has actually moved into the position requested. The locking bed shown is for a General Railway Signal (GRS) power interlocking machine.

Relay interlocking

Interlocking effected purely electrically (sometimes referred to as "all-electric") consist of complex circuitry made up of relays in an arrangement of relay logic that ascertain the state or position of each signal appliance (Figure 29c). As appliances are operated, their change of position opens some circuits that lock out other appliances that would conflict with the new position. Similarly, other circuits are closed when the appliances they control become safe to operate. Equipment used for railroad signalling tends to be expensive because of its specialized nature and fail-safe design. Interlocking operated solely by electrical circuitry may be operated locally or remotely. Furthermore, such an interlocking may be designed to operate without a human operator. These arrangements are termed automatic interlocking, and the approach of a train sets its own route automatically, provided no conflicting movements are in progress.

Electronic interlocking

Modern interlocking (those installed since the late 1980s) are generally solid state, where the wired networks of relays are replaced by software logic running on special-

purpose control hardware. The fact that the logic is implemented by software rather than hard-wired circuitry greatly facilitates the ability to make modifications when needed by reprogramming rather than rewiring (Figure 29d). Regardless of the technology used, interlockings are designed to ensure that no operation can be performed unless all prerequisites have been satisfied. Solid State Interlocking (SSI) is the brand name of the first generation microprocessor-based interlocking developed in the 1980s by British Rail, GEC-General Signal and Westinghouse Signals Ltd in the UK. Second generation processor-based interlocking are known by the term "Computer Based Interlocking" (CBI). They are based on a verification process which involves three computers, and a certain locking action is applied to field movable elements only if at least two out of the three computers (the so-called 2oo3 logic) identify the same action to perform. This logic in fact meets the fail-safe principle required by railway interlocking operations.

Chapter 3. An Overview on simulation models of Railway Systems and their applications in practice.

3.1. Introduction

As illustrated in the previous chapter, railway networks can be classified as systems characterized by a high degree of complexity due to the large amount of complex interactions existing among its components. Such complexity, however prevents system behaviour from being accurately described by closed-form analytical solution, and it is necessary to rely on apposite simulation techniques. Simulation models of railway systems are in fact largely spread over academic and professional practice in supporting decisional activities of planning and design phases, as well as real-time traffic management. These simulation models can be distinguished according to the level of detail through which they represent railway network (macroscopic, mesoscopic, microscopic), to the type of assumption made for the occurrence of events (stochastic or deterministic) and to the kind of the processing technique of the events (synchronous, asynchronous). In literature a wide application of railway simulation models can be observed to face and solve different aspects of planning and design phases. Simulation approaches have been in fact adopted for supporting stability analysis (*Nordeen, 1996, Middlekoop & Bouwman, 2002*) or the robust design of service timetables (*Middlekoop & Bouwman, 2000*). Other relevant applications have been observed instead for improving the block section layout with respect to system capacity (*Gill & Goodmann 1992, Chang & Du, 1998, 1999*), or with respect to energy consumption (*Ke et al. 2009*). Many applications are also considered in real-time management of railway traffic to solve conflicts and/or efficiently rescheduling service after a disruption.

In this chapter therefore after a brief description of the different kinds of railway simulation models, an overview on the different models developed throughout the world will be given. Successively some fundamental applications addressed to answer some issues proper of both practical and academic areas will be described.

3.2. Network Modelling

Railway infrastructure can be modelled relying on structures derived from graph theory. In fact Graph Models (*Hauptmann, 2000*), (*Ratke, Watson 2007*) have the main advantage of being very flexible and able to describe also complex railway

infrastructure in an efficient mathematical model. Moreover, such models allow modular and redundancy free storage and handling of the data in computer models. The real existing railway infrastructure (tracks, signalling systems, or attributes like gradient, radius, etc.) is separated into pieces (links). Then such links are bounded by nodes which just join the links.

To better understand railway infrastructure modelling, it is necessary to specify how nodes and links of a graph model represent the several elements of a real network. In particular:

- *Nodes* (also called in mathematical terms “vertexes”) are a representation of an arbitrary location of a punctual elements (e.g. signals or switches for microscopic models or stations for macroscopic models)
- *Links* (mathematically called “edges”) can be defined instead as the connection between two nodes (therefore can coincide with block sections for microscopic models or inter-station tracks in macroscopic models)

Infrastructure models are generally represented by the so- called “valued” graphs where a certain weight (usually representing a cost) is assigned to links, nodes or both. The general mathematical formulation for a valued graph is reported in equation (23):

$$G \equiv (N, L, c), \quad \text{with } c(l) \geq 0; \quad \forall l \in L \quad (23)$$

where N and L are respectively the set of all nodes and links belonging to the considered graph, while c is the weight function relative to graph links.

According to the number of links which connect two consecutive nodes, graphs can be classified as:

- *Directed*, if two consecutive nodes are linked with at least one link which is generally orientated and its direction is indicated with an arrow.
- *Simple*, if the graph G does not contain parallel links or loops.
- *Connected*, if for any two nodes of G , links exist connecting the nodes.

Moreover, employing graphs as mathematical models, it is possible to face different railway related problems, which can be translated in mathematical terms and

successively stored and processed by computer algorithms: *a)* Concrete railway infrastructure, *b)* Abstract dependencies (rules or flows).

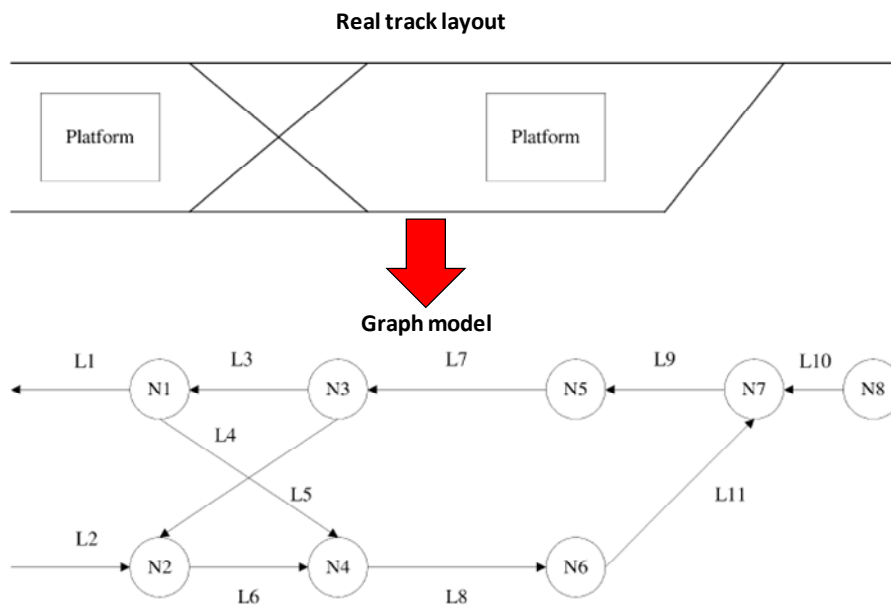


Figure 30. Representation of a real railway track layout as a graph model.

Figure 30 shows the example of a real track layout composed of two cross rails for inverting train direction, and the corresponding representation as a graph model. As can be seen the reported graph is directed (since links have a specific direction) and simple (because no loops are contained), but it is not connected since not all couples of nodes are linked together.

However, for various optimization problems and their depiction in graphs, very efficient mathematical algorithms and heuristics exist. For the modelling of railway infrastructure, the following problems are among the most important:

- Calculation of the shortest paths in graph networks between two nodes according to different criteria (shortest path problem). For example the Dijkstra algorithm is usually used to solve this kind of problem (*Dijkstra*, 1959).
- Calculation of maximum or minimum flows in graph-networks also considering the respect of capacity constraints of links and nodes, independently from time. To these purposes algorithms such as the Edmonds and Karp one, could be employed (*Schumacher* 2004), (*Radtke* 2005).

Furthermore, other railway problems like for example the allocation of vehicles can be also described relying on graph theory but using more complex approaches such as the multi-commodity flow, maximum concurrent flows, dynamic flows or even a combination with heuristic algorithms (e.g. simulated annealing, genetic algorithm), as presented by Kettner (2005) and Schumacher (2004).

As said before, the attributes of a railway infrastructure modelled in a graph are assigned to links and nodes, in fact typical attributes for nodes are geographical position of point elements (e.g. station positions) as well as the kind of element depicted by that node (e.g. if the node represent a station, a signal or a block section joint). For the attribution of links instead two main approaches can be distinguished.

- **Link-orientated models**, where each link contains all relevant information relative to tracks such as speed limits, gradients, curvature radii, electrification, direction, etc.
- **Node-orientated models**, where instead links do not contain any information about the track, but special node types indicate the change of attributes (e.g. speed by a speed change node, gradient by a gradient change node). Moreover each node has positional coordinates and the length of a link can be calculated as the difference between the positions of the end and the start nodes.

Both approaches have advantages and disadvantages. The redundant assignment of all railway infrastructure attributes to each single link in a link-orientated model can be seen as a disadvantage. Wasting storage capacity and implying problems handling changes to those attributes. If a user needs to change a certain attribute in a defined area, all links of that area have to be considered. In some cases, links have to be separated and a new node created in order to obtain a modified assignment of attributes.

Anyway modern infrastructure editors provide efficient support with easy to use functions to overcome the data-handling problem. Moreover, thanks to the development of computer industry the amount of data to be stored in the RAM is no longer a critical issue, since the possibility of coupling together several RAM items.

A node-orientated attribution model, has the advantage of storing no-redundant data. For instance, to introduce a speed change, a node of the type “speed-node” has to be set at the desired position, or an existing node has to be changed. On the other side, node-

orientated models require complicated algorithms to calculate all the attributes on a certain link of the network. This task can be moreover very complex when considering the kilometre markings in railway networks, which are often not well defined due to historical reasons. Changes of the kilometre marking direction (going up, going down) and the handling of wrong kilometre values require significant efforts. In general, link-orientated models do not need such functionality.

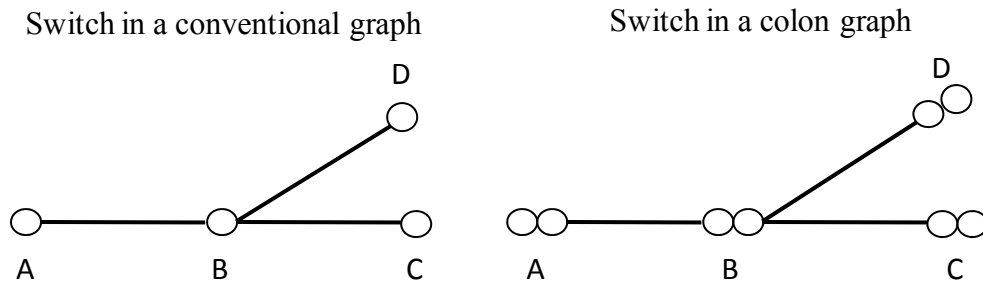


Figure 31. Representation of a switch in a conventional and in a colon graph.

An alternative approach to node modelling has been proposed to represent railway networks. Such approach is called colon graph and it assures an easy way to the right utilisation of points during railway operation. Figure 31 shows in fact the difference between the depiction of a switch element in a classical graph model and a colon graph representation. As can be clearly seen within the colon graph convention each node is doubled, in order to establish also the direction according to a certain path must be followed. According to such convention in fact a path in the colon graph must always satisfy the sequence: “node-node-link-node-node-link-node-node”. Therefore, for the example reported in Figure 31, if a train is starting its path from node D, it can follow only the path D-B-A and not the path D-B-C, since this one does not meet the sequence imposed by the colon graph convention.

However, according to the level of detail through which railway network is represented, it is generally possible to distinguish amongst three different kinds of models: *macroscopic*, *mesoscopic* and *microscopic* models. From Figure 32 it is immediate to understand the strong decrease in information detail when passing from a microscopic to a macroscopic representation. In fact a microscopic station or junction becomes a simple node of the graph within a macroscopic model, losing therefore a consistent amount of infrastructure information. Anyway each one of these models will be described in detail within the following sections.

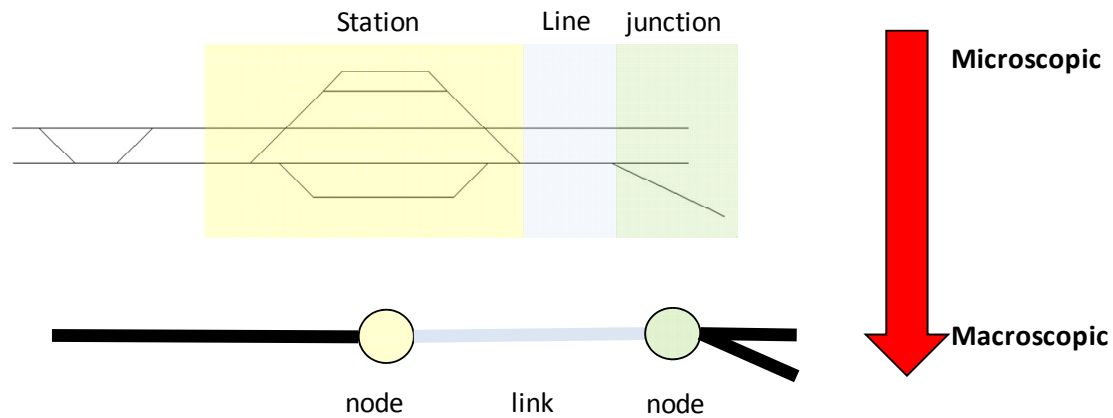


Figure 32. Representation of a network section at a microscopic and macroscopic level.

3.2.1. Macroscopic models

Macroscopic models are usually preferred to support long term planning tasks or special routing problems, since they represent network infrastructure only at high level of abstraction and for such reason they can be only employed when track information are known in an approximate way. It is clear that macroscopic model contain far less links and nodes with respect of a microscopic model. For instance the entire German Railway Infrastructure which could be represented at microscopic level with about 850000 links and 830000 nodes, can be described instead at macroscopic level by a graph containing only 10000 links and 9500 nodes.

As said before a node of a macroscopic model can represent a station or a junction of the real network, independently from the complexity of the station or the junction itself (therefore a node can depict both a stop station that a complex terminal station).

Macroscopic data can be entered manually from various sources: even public (internet) sources are sufficient in some cases. If a microscopic infrastructure database is available, a direct transfer from the microscopic to the macroscopic model is preferable.

A typical macroscopic node would contain the following information:

- *Geographical attributes* (coordinates, names),
- *Type of node* (station, joint, shunting yard, etc.).

Macroscopic links instead may hold information relative to:

- *Length*

- *Type of line* (high speed, passenger, freight, etc.)
- *Number of tracks*
- *Train availability*
- *Average running time*
- *Average capacity* (for example according to UIC 406 code)

One main application of macroscopic network models for long term planning activities is for example the search for train paths without time restrictions in networks. Introducing new trains during traffic assignment, finding a path for a locomotive for empty runs, or rerouting of freight trains always requires the search for a feasible path in a railway network. In practice this is a two-step process (Sewczyk 2007). The first search on a macroscopic network does not consider all individual tracks on lines or stations but determines a sequence of stations for the train trip. Some fundamental criteria considered are axle load, electrification or other operational rules like stopping patterns for passenger trains or preferred routes for high-speed trains. In complex networks, calculation time can be reduced. After finding an initial solution (e.g. applying the Dijkstra algorithm) the “fine tuning” of the train path can be finished on a microscopic network later, working out all tracks on lines and through stations.

It is possible to control the routing process even without the exact depiction of single tracks in stations using the method of the inverse graph (Kettner 2005). This problem may occur in terminal stations where the inverse graph allows the correct modelling of feasible paths along the lines and the terminal station.

Macroscopic models are therefore not suitable for all detailed planning tasks like the calculation of running times or conflict detection. In general, these models do not contain information about the exact maximum speed and track gradient or restrictions due to the signalling system. Therefore, it is not possible to calculate a correct running time. The approach of using an “average” gradient or speed in macroscopic models will mislead a user, and running time calculations based on these data should not be employed in serious researches or consulting works.

Several macroscopic models have been developed by both consulting firms and academic bodies for commercial or research purposes, and here two of these models

implemented within European area: NEMO and SIMONE, will be described as examples.

The NEMO model

The railway macroscopic model for simulating railway operation NEMO (Network Evaluation Model by ÖBB) has been developed by the IVE (Institute of Transport Railway Construction and Operation) at the University of Hannover, in collaboration with the Austrian Federal Railways (ÖBB). It is a strategic planning tool for evaluation of infrastructure and operational measures in railway systems. It is based on a macroscopic network modelled as a link-orientated graph of the system, where nodes represent point elements like stations, junctions, shunting yard, and links depict inter-station tracks. Substantially, it models the interactions between railway system (intended as network infrastructure and railway operation) and transport demand (Kettner *et al.* 2001, 2002, 2003).

In particular NEMO provides a way to efficiently evaluate different production and infrastructure scenarios both for passenger and freight traffic. Transport supply can be moreover optimized based on demand and capacity. Operational bottlenecks can also be detected. Thus the model supports efficient use of investment funds, since also economic evaluation of planning scenarios can be performed. The model is composed of 4 modules which respectively interact as shown in Figure 33. Specifically:

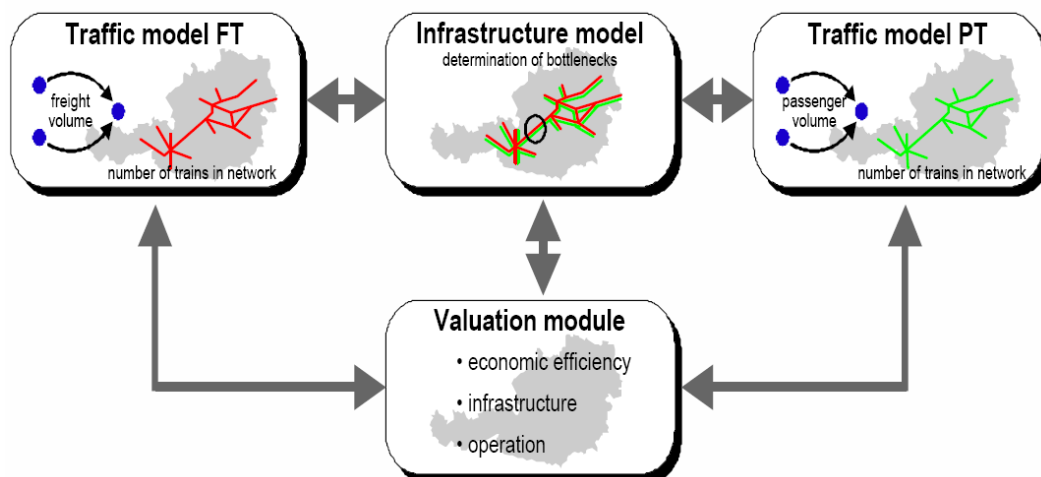


Figure 33. Structure of NEMO model (Kettner *et al.* 2003)

- The *Infrastructure module* describes the railway network as a link-orientated graph where nodes represent all point network elements (e.g. stations, junctions, etc.) and

links depict inter-station tracks. The graph model contains moreover the access points (centroids) for passenger and freight traffic, and links take into account the capacity of the corresponding inter-station section as well as the average travel time for the different train types, to allow a correct traffic assignment. In Figure 34 a macroscopic representation of a railway network as used in NEMO is reported.

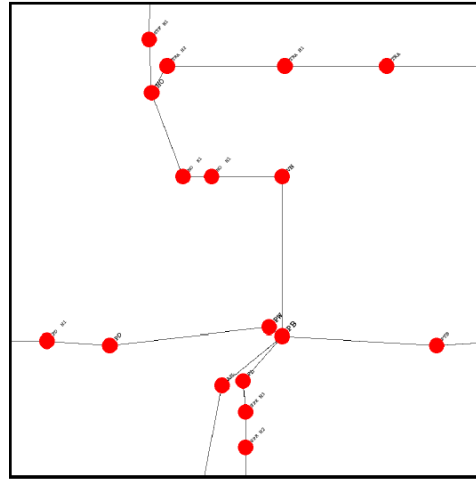


Figure 34. Macroscopic network representation by NEMO model (Kettner et al. 2003)

- *Traffic module* is addressed to model train operations on the infrastructure, therefore requires as input data a certain number of train models, which provide for each train category: running times for each inter-station track, minimum headways, and other information about that train category (e.g. passenger or freight train, priority, average speed). Such parameters in fact are used within the assignment model and therefore are fundamental to calculate the occupation times of each infrastructure section, and therefore to identify capacity bottlenecks. Moreover parameters of train model strongly depend from microscopic train characteristics like the type of traction unit mounted on the rail vehicle, the weight, the number of wagons, etc. Furthermore, since each train model is assigned to the links of the infrastructure network graph which belong to its path, it is possible for instance, to model a local separation of high-speed passenger trains and low-speed freight trains in the model.
- *Evaluation module* contains different performance indexes which are of key interest for the objectives of the investigation, consenting therefore the estimation of effects induced by a certain design solution on system performances. Index values in fact are determined through processing simulation outputs returned by the NEMO model. In particular each planning scenario (with a certain infrastructure layout and

operational schedule) can be evaluated measuring both economic and system performances (e.g. capacity bottlenecks, average train delays).

The SIMONE model

The SIMONE model has been developed by the Dutch ProRail (*Middelkoop, Bowmann 2001*) to simulate and analyze complex and large scale railway networks. In particular the main purposes of this model concerns: (1) the assessment of timetables robustness; (2) the determination of network stability: analyzing causes and effects of delays; (3) the improvement of timetables, by determining the relations between design standards and robustness of the timetable; (4) the detection and quantification of bottlenecks in a train network; (5) the quantification of delays for different layouts of railway infrastructures. Moreover a strong feature of SIMONE is the ability to automatically generate ready-to-use network simulation models from databases. The architecture of this macroscopic model is based on eight different modules as illustrated in Figure 35.

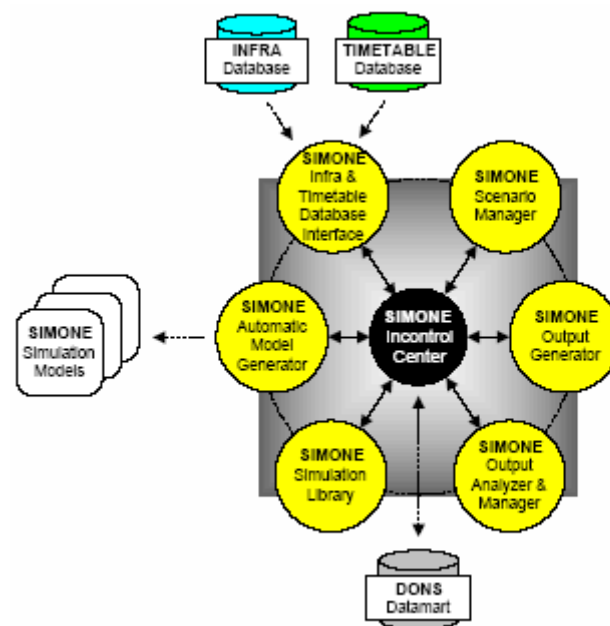


Figure 35. Architecture of the SIMONE macroscopic model (Middelkoop, Bowman 2001).

In particular the main SIMONE modules are:

- The *Incontrol Centre* is the core of the simulation environment from where everything is controlled. Information is stored in the Oracle database of the Railed (capacity manager of the Dutch rail infrastructure) system for later retrieval. This information ranges from experiment setups, to user annotations and to the output experiments.

- The *Simulation Library* is a collection of six simulation building blocks, which takes as input data all physical characteristics of network (e.g. positions of stations and junctions, length and capacity of each inter-station tracks), of trains (e.g. travel times for each kind of train category, type of train), and operational timetable (e.g. headways, departure, arrival from/at station or junctions, disturbances to service).
- The *Infrastructure and Timetable Interface* is addressed to generate simulation models based on information on timetables and infrastructure contained in databases. In particular such module is designed for interfacing SIMONE with the Database of the Dutch railway systems, the so-called DONS. Then on the basis of the model description of both infrastructure and traffic demand, provided by DONS, a cyclic timetable is generated.
- The *Scenario Manager* contains the GUI (graphical User Interface) of the network model. Through this module it is possible to have a graphical representation of the constructed model and visualize graphical outputs of the simulation like for example train running along the network (coloured in different way according to the category to which it belongs to). Figure 36 shows the representation of the Dutch railway network as given by SIMONE.

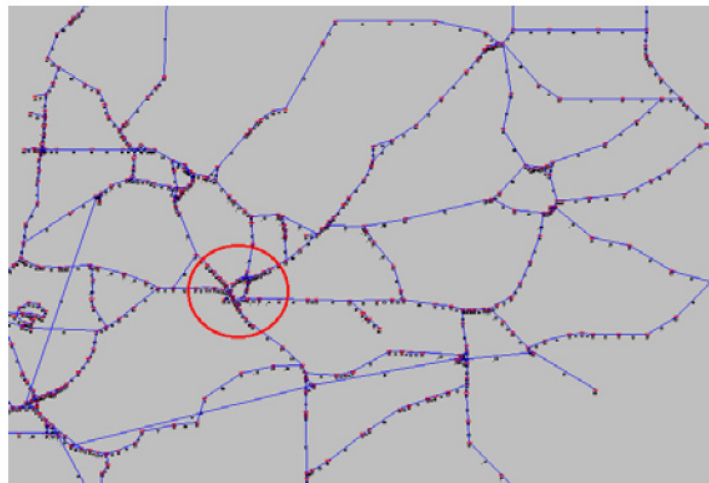


Figure 36. Macroscopic model of the Dutch railway network as represented in SIMONE (Middelkoop, Bowman, 2003).

- The *Output Generator* module is the responsible for the provision of simulation outputs. In particular such output can be gathered for the whole network or only for smaller areas such as stations or junctions. This module in fact collects all simulation output and automatically creates reports in both numerical and graphical

formats. For example Figure 37 shows how train delays can be graphically represented on the network by the SIMONE model.

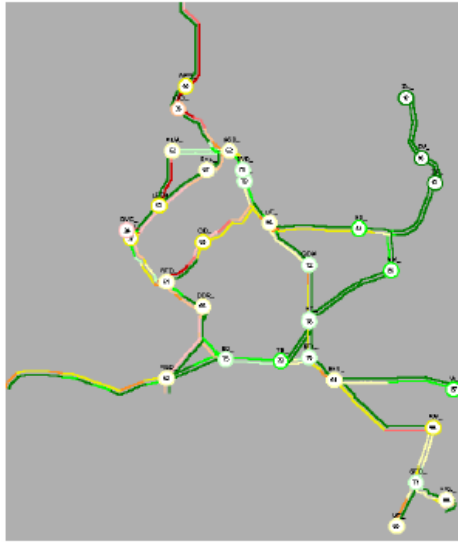


Figure 37. Graphical representation of the Train Delay output (red: higher delay, yellow: lower delays) as returned by SIMONE model (Middelkoop, Bowman, 2003).

3.2.2. Mesoscopic models

The mesoscopic infrastructure model is placed in between the macroscopic and the microscopic representation of railway network. This kind of model can be generated for specific tasks such as the simplified railway “simulation” of complex networks to answer some strategic and tactical questions. This kind of models are classifiable as “multi-scale” models, since they contain both areas modelled on a microscopic level (e.g. signalling system, layout of marshalling yards) and areas modelled instead on a macroscopic level (e.g. stations). The advantage of such models is however relative to the minimization of effort for modelling aspects of a complex problem that are of insignificant relevance for the overall outcome of the investigation (e.g. the individual operation of shunting yard or vehicle depots for the planning of a network wide timetable).

The difference between a macroscopic and a mesoscopic network can be seen in Figure 38. It is clear that with respect to macroscopic model it is possible to represent at a microscopic level the signalling system implemented on the track and represent therefore the block section layout with a certain accuracy. However the correct location

of signals and block length can only be modelled in microscopic models, and therefore reliable capacity figures are not achievable with mesoscopic infrastructure models.

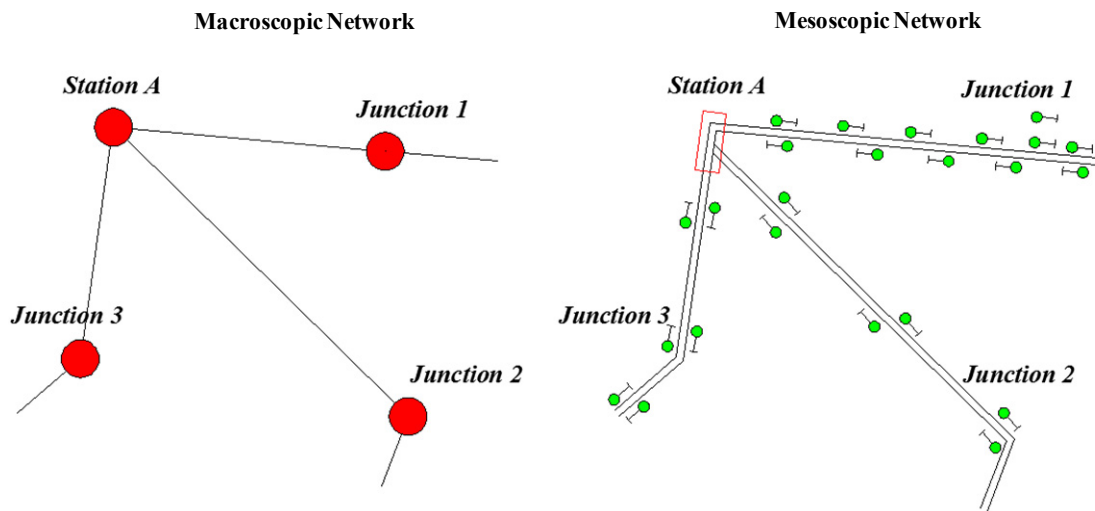


Figure 38. Comparison between a macroscopic and a mesoscopic network model.

Furthermore, another element of this approach that must be noted is that with respect to macroscopic models, it is also possible to have a more detailed representation of point elements such as stations (here called in fact “mesoscopic station”) since the individual station tracks and possible paths from the lines to the station tracks, can be modelled. Anyway, sometimes it might be necessary to increase the level of detail of multi scaled infrastructure models for specific applications. This can be obtained without any problems only if correct microscopic data are available.

Nowadays, at the best of our knowledge, there are not commercial mesoscopic models for simulating railway traffic, but several mesoscopic models developed to meet pure research requirements can be observed in literature.

In particular here a mesoscopic model for analyzing and evaluating freight train operation (Marinov, Viegas, 2011) is described as an example of this kind of models.

A mesoscopic model for simulating freight train operations

As said before, the mesoscopic model reported in this section is not a commercial model, but it has been developed by Marinov and Viegas (2011) to satisfy research and operational objectives such as the evaluation of how different freight train deviations from schedules affect performances of a complex flat-shunted yard in a railway

network. In particular this is an “event-driven” model implemented by using an event based simulation package called SIMUL 8. Such model adopts a decomposition approach, in the sense that the different infrastructure components that compose railway system (rail yards, rail terminals, rail line: single and double track, junctions and passenger stations) are modelled separately, in order to decrease the complexity of the whole system and increase the accuracy in the estimation of performances. However such accuracy strongly depends on the way into which this decomposition is performed and above all on how the interactions amongst these components are modelled. According to this assumption therefore, it is possible to say that the railway network is here modelled as a queuing network, where all components are interconnected queuing systems that interact and influence one another, so that the global impact of individual and all freight train operations in a network is captured.

Specifically this mesoscopic model simulates operations following these rules:

- *Attributes.* The simulation is built up using a set of Work Centres (i.e. servers) and Storage Areas (i.e. buffer or queues). The Work Centres (rail line, rail yards, etc.) serve work items (here trains) which are the “clients” of the system. In our case the Work Centres are used to replicate the operating processes with freight trains, or in other words this is where a freight train is served by a component of the rail network. Each Work Centre is characterized with inbound traffic, service pattern and outbound traffic. The inbound traffic is the number of freight trains that require for service in the Work Centre. The service pattern follows a particular distribution which means information is obtained by observations, real data collection and statistical analysis. To generate the work items (i.e. trains) an attribute called Work Entry Point is employed, which is characterized with arrival pattern. The arrival pattern may be subordinated to a theoretical or an empirical distribution, as well as a time-dependent distribution. The outbound traffic is instead the outcome of the Work Centre and is routed to other Work Centres and Storage Areas in a variety of ways (e.g. to a subsequent Work Centre, to a Storage Area waiting for next operation, or to leave the simulation when all operations have been completed). The Storage Areas are attributes used to replicate where the freight trains are held while waiting to be processed by a given component of the rail network. All Storage Areas are controlled by their capacity.

When a train leaves the network its service is assumed as terminated, In SIMUL 8 this event is replicated by an attribute called Work Exit Point.

- *Arrival Pattern*: time-dependent distribution. This kind of arrival pattern is very useful when modelling systems that do not reach steady state as a railway system where trains have to run on strict fixed schedules. Train arrivals in fact can vary over a regular cycle and the system deals with predictable variability, predictable queues and quasi-steady regimes of work. In predictable variability there are busy periods in which queues tend to build up, and quiet periods where instead queues tend to reduce.
- *Routing*. The routing of trains (work items) in the network is specified by “Work Flow Arrows”, which is to indicate the paths (from Work Entry points to Storage Areas or to Work Centres, etc.) for the trains moving through each component in which the railway network has been decomposed. An important feature here is that when there is no Storage Area (i.e. there is no buffer to queue) between two successive network components, but there is a direct link between them, by default the freight trains do not move from the first subsystem to the second until the second subsystem is ready to start processing them.
- *Measures of system performance (MOPs)*. Simulation outputs considered by such model are: total number of trains processed by a given Work Centre (i.e. the consumed capacity for each network component), the number of trains in a given Storage Area (i.e. the length of a queue), queuing (waiting time) per train on average for the simulation period. Moreover results of simulation experiments, and therefore the values of each MOP are considered as the average on a certain number of scenario replications, in order to reduce the stochasticity.

Figure 39 illustrates how a rail marshalling yard (whose microscopic representation is reported in Figure 39a) can be modelled at a mesoscopic level (Figure 39b), by using the described “event-based” model implemented in SIMUL 8.

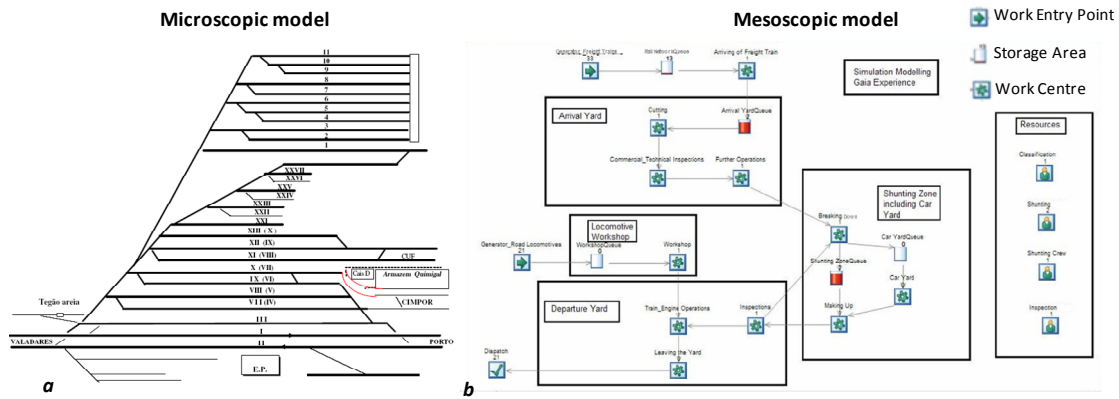


Figure 39. Mesoscopic model of a rail yard implemented in SIMUL 8 (Marinov & Viegas, 2011).

3.2.3. Microscopic Models

Microscopic models depict railway networks at a high level of detail, representing infrastructure with a detailed node-link model. The microscopic infrastructure model in fact combines detailed track information like speed limits, gradients, curvature radii of rail sections, with signalling system (signals, block sections, release points) and some operational information like routes, alternative platforms and timing points. Every single change in one of the above considered attributes requires a new node splitting an existing link and generating a new one. According to the type of graph used to model the network, information can be assigned to nodes (in a node-orientated model) or to links (in a link-orientated model). Therefore, given the granularity of microscopic models, they are not suitable but also mandatory for exact running time calculation, timetable construction and simulation, conflict detection and resolution. A typical microscopic infrastructure model contains all tracks on both lines and stations. In Figure 40a a microscopic representation of a railway section is represented. It is possible to see that also signalling layout is depicted in detail. In fact a block section, which is delimited by two consecutive signals, is formed by several links. The block section itself has a starting and an ending node, both of these nodes belonging to existing individual links. Therefore, in the model no additional nodes to describe the block are necessary.

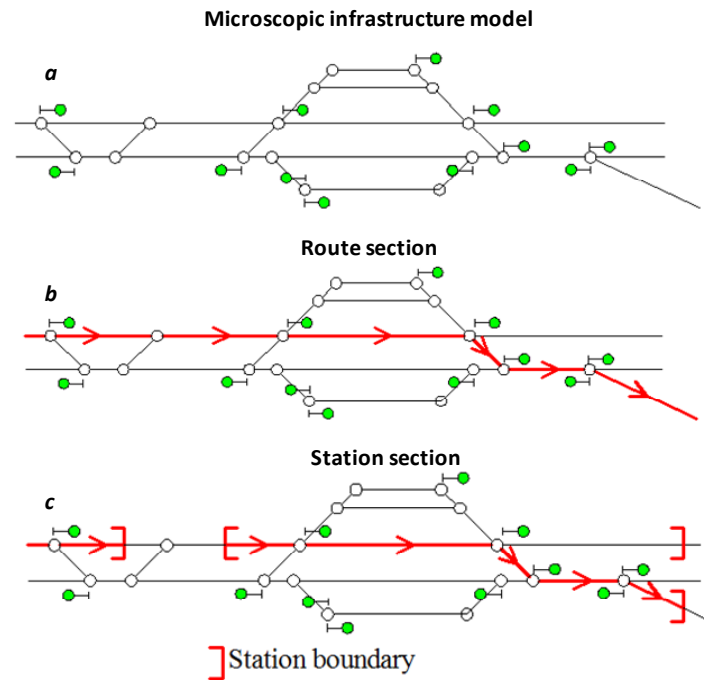


Figure 40. Microscopic representation of a railway section (a), route section (b) and station section (c) to determine train paths on the network.

Several block section can be abstracted to route sections (Figure 40b), which are sequences of links guiding the trains through the station. They are provided by the interlocking system in the respective station. Trains are assigned to route sections which can be used for specifying the correct train path through each station, finding a correct train path during timetable construction or railway simulation, or assigning signal protected paths into a complex station to a platform.

Therefore a route section can only be created if this section is feasible from the operational point of view, for example a route section, which is supposed to be used by electric locomotives, must not be created in a model if some links are not electrified. However this simple approach can be used to transfer technical requirements or operational rules into the model.

Slightly different modelling approaches use station sections for the definition of train paths (Figure 40c). Station sections are sequences of connected links within a single station, starting and ending at so-called station boundaries. They are provided by the interlocking system in the respective station. Trains are assigned to route sections for specifying the correct train path through each station. This concept has the advantage that complex train movements in stations can clearly be described in the model.

Since microscopic models describe in detail each component of railway system they need to be fed by a large amount of input data such as: length, speed limits, and radii of links, signalling system features (type, overlaps, block sections, release points, track circuits), vehicle characteristics (weight, characteristic “speed-tractive effort” curve, number of wagons), interlocking techniques, electrification, etc.

As can be easily understood, the large number of input data required consents a very accurate modelling of railway operations but on the other hand, it makes this kind of models be inefficient from a computational point of view when simulating large-scale networks or when supporting analyses which needs a consistent number of simulations (e.g. probabilistic analyses, black-box optimization problems).

However, such models are used throughout the world for supporting designing activities of both infrastructure elements and operational rules. From the experience gained, it is clear that to obtain reliable results the interrelation of the infrastructure, the rolling stock and the timetable should be assessed on a microscopic network level, which includes lines and stations. The complexity considered by such a system can include various signalling systems, different vehicle dynamics and train scheduling rules as well as additional operational conditions. All these issues are strongly dependant on the modelling of the railway infrastructure. Furthermore working in complex infrastructure networks increasingly becomes a standard approach. For example, the identification and evaluation of reliable capacity indicators for railway lines requires the inclusion of at least the adjacent junctions. It became apparent during the practical application of timetable construction and simulation methods that all branch lines of those junctions should be considered in the model as well, to ensure that the simulation could incorporate the impact of trains arriving at the junctions from those branch lines.

Several academic and commercial microscopic models have been developed around the world to satisfy both research and practical objectives. In particular it is worth mentioning commercial microscopic models like *OpenTrack*[®] developed by the ETH-Zurich, *RailSys*[®] realized at the Leibnitz Universitat of Hannover, and *RAILSIM*[®] distributed in the USA by SISTRA and RTC in collaboration with Berkley Simulation Software. In the following sections the models *OpenTrack*[®] and *RailSys*[®] will be described in further detail.

The OpenTrack model

OpenTrack[®] is a microscopic model for simulating railway operations developed by A. Nash and D. Huerlimann at the ETH of Zurich, Switzerland. This is a time-driven stochastic model, able to simulate at the same time several trains (multi-train simulator), and it is basically composed of three modules according to which its input data are administered. As can be seen by Figure 41 these modules are:

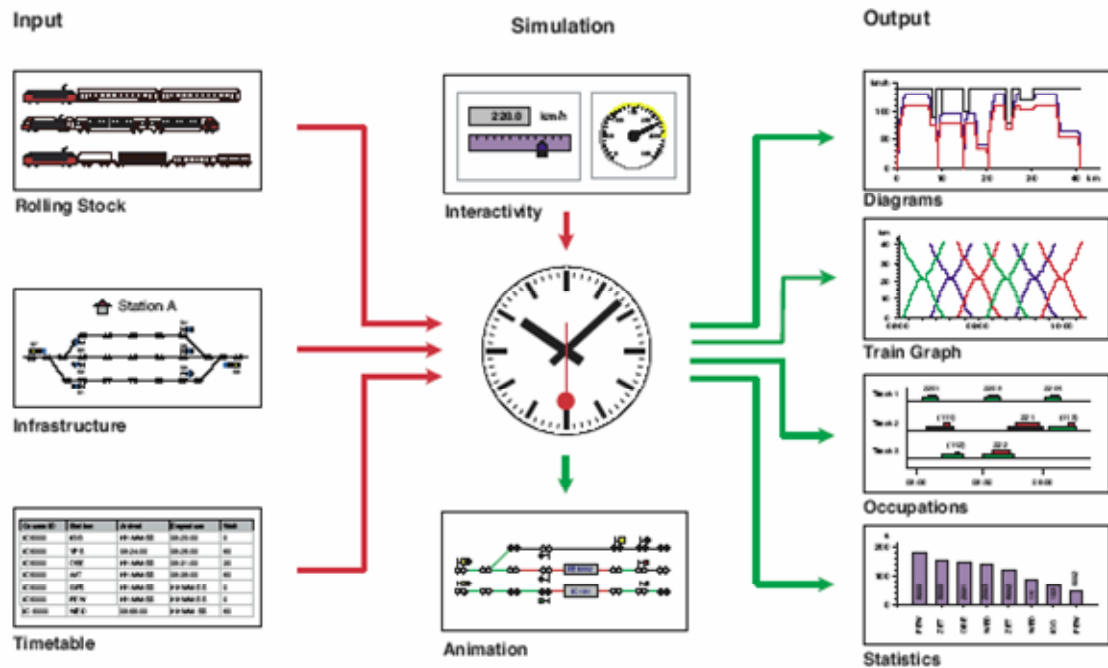


Figure 41. Architecture of the OpenTrack model (Nash, Huerlimann 2004).

- *Rolling Stock Module.* In this module all physical and mechanical characteristics of rail vehicles must be specified for each train category considered within the simulation (e.g. Intercity, Regional trains, Urban trains, etc.). In particular parameters such as weight, length, number of wagons, “speed-tractive effort” characteristic curve of the locomotive, etc. must be specified as inputs in this module. Moreover a module is included in such module to calculate train energy consumed during its operations on the network.
- *Infrastructure Module.* Here railway network is modelled as a link-orientated graph, in which all physical information of the infrastructure is considered as attributes of links, while nodes represent positions of stations, signals, switches and other point components. Therefore input parameters like gradients, radii, speed limits, electrification of rail tracks as well as positions of stations and switches, must be all

specified in this module. Moreover, it is necessary also to determine the kind of signalling system implemented, the length of block sections through positioning their delimiting signals on graph nodes, and the type of interlocking systems which regulate train movements within station or junction areas.

- *Timetable module.* This module considers as input data scheduled arrival/departure times of train runs at/from stations. It is also possible to specify only the dwell time of each train run for each considered station. Moreover such dwell times can be assumed both as deterministic and random variables (distributed according different probability density functions), and in the latter case it is possible therefore to take into account for stochastic disturbances which affect real operations. Furthermore in this module it is also possible to set a certain percentage of delayed train runs at some stations.

Since OpenTrack is a “time-driven” model, the simulation of train movements is realized by calculating for each time step train speeds and positions through the integration of Newton’s motion formula. Indeed, train motion parameters (speed and position) also must respect at each time step the aspects of signals ahead which change according to the position of preceding trains occupying successive block sections.

Output data returned by an OpenTrack simulation are instead:

- *Train motion diagrams* (speed-distance, speed-time, distance-time trajectories)
- *Occupation times* of rail sections (in both numerical and graphical format)
- *Statistics*, such as percentage of delayed trains at a certain station, overall train punctuality (fixing a certain delay threshold).
- *Energy consumption diagrams* (electrical or mechanical power-time diagrams, electrical or mechanical energy-space diagrams).

In Figure 42 the graphical representation returned by OpenTrack for a simple railway network is shown. Furthermore, some of the aforementioned outputs are illustrated within their graphical format.

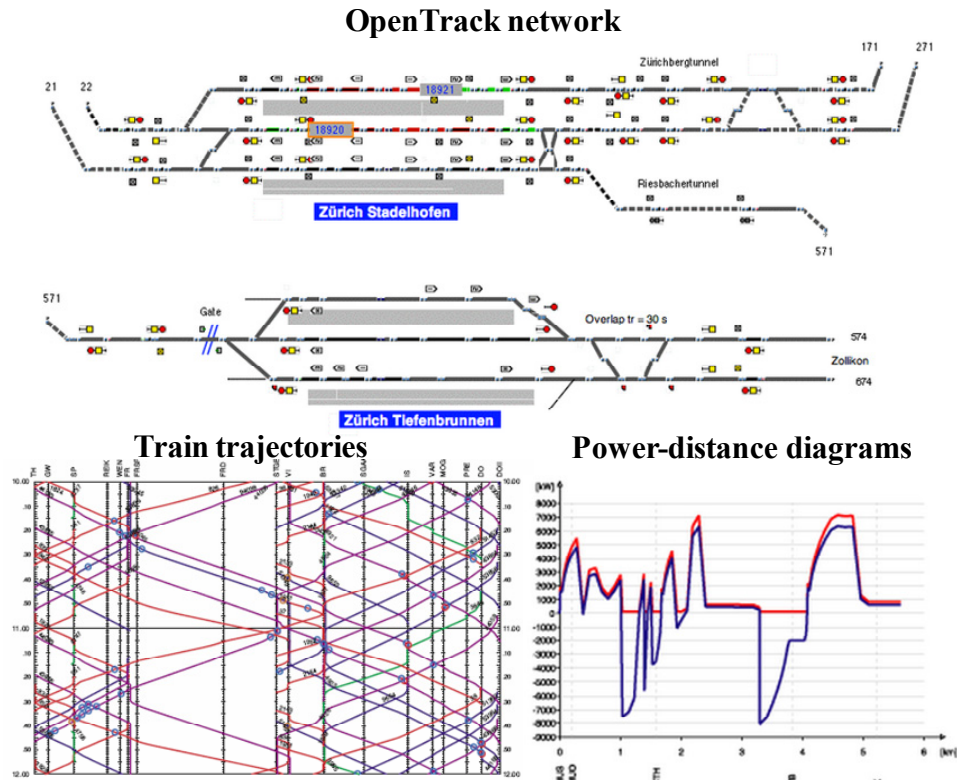


Figure 42. OpenTrack graphical representation of a railway section, of train trajectories and power consumption.

The RailSys model

RailSys[®] is a time driven microscopic model for simulating and planning railway operations. It has been realized initially by the Leibnitz Universitat of Hannover, and then further developed and distributed in cooperation with Rail Management Consultants (Siefer, Radtke, 2004). Also this model is structured according to different modules, some of which are addressed to administer input data, while other are dedicated to elaborate simulation outputs for evaluating system performances under a certain simulation scenario. Specifically, as can be seen by Figure 43, the main modules which compose such a model are:

- *Infrastructure manager*. This module administers all microscopic infrastructure information which just constitutes the main input data. Therefore, station positions, switch and signal locations must be all assigned to the nodes of the link-oriented graph through which the railway network is modelled. Graph links instead require the specification of track gradients, speed limits, curvature radii, etc. Moreover here

it is also possible to define the type and the layout (e.g. block section lengths) of the signalling system implemented on the network.

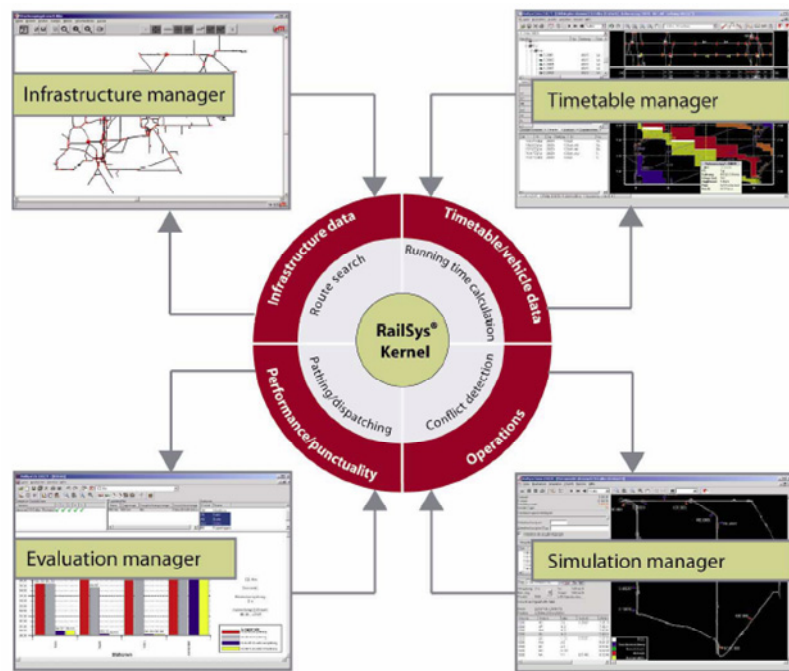


Figure 43. Architecture of the RailSys model (Siefer, Radtke, 2005).

- Timetable manager.* Input data required by such a module are relative to train departure/arrival times at stations or particular junctions, or also station dwell times can be specified. In particular when running synchronous simulation of the network, such parameters can be considered as constant or random variables, in order to take into account for the stochastic disturbance occurring during service. Furthermore this module includes a part for conflict-free timetable construction (Radtke et al. 2004) making use of train minimum running times returned by a deterministic simulation of train runs along the considered network. Specifically, after a first time driven simulation of operations respecting a certain initial timetable, both train travel times and train conflicts are determined as simulation outputs (this task is performed by a sub-module called DYNAMIS which considers all physical-mechanical features of rail vehicles to determine correct running times). Then such results are transferred to the timetable construction phase (done by a sub-module called Simu++) which finds and solves train conflicts due to blocking times overlaps, returning therefore a conflict-free timetable. In Figure 44 the architecture of the RailSys timetable environment is reported.

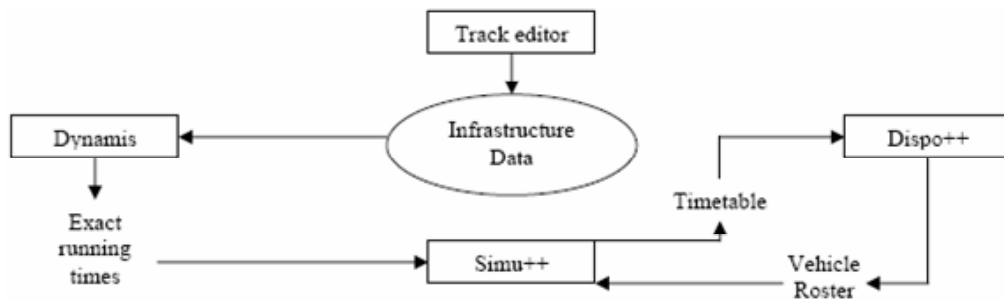


Figure 44. Architecture of RailSys planning environment (Bendfeldt et al., 2000).

- *Simulation Manager.* In this module (and in particular the DYNAMIS sub-module) it is necessary to specify all physical and mechanical characteristics of rail vehicles (e.g. weight, length, number of wagons, “speed-tractive effort” curve of locomotive, etc.), in order to simulate train operations along the network. In particular, as said for the previous module, simulation is divided in two parts in RailSys: the first part regards the simulation of a timetable to check conflicts over the network that remain unsolved, while the second part is an operational stochastic simulation, introducing randomly additional delays to trains. The aims of this second type of simulation are relative to the check and the quantification of timetable quality, as well as the estimation of effects induced on performances by disturbances.
- *Evaluation Manager.* Results returned by the simulation module are transferred and then elaborated by such module. In particular model outputs can be used for statistical determination of train performances (e.g. arrival delays, number of trains which are delayed over a certain threshold, etc.), train trajectories, occupation times, speed-time diagrams etc.

Figure 45 illustrates the graphical representation of a railway network modelled in RailSys, and several graphical outputs returned by both the Timetable manager sub-components (Simu++) for constructing a conflict-free timetable and the Evaluation manager relative to train delays statistics.

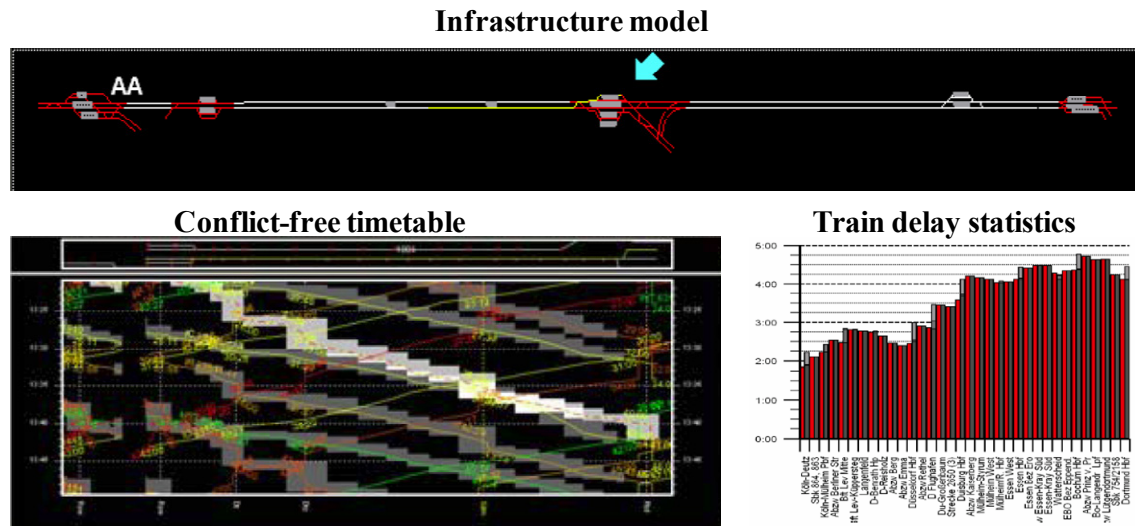


Figure 45. RailSys graphical representation of infrastructure, designed timetable and train delay statistics.

3.2.4. Hybrid models

The term Hybrid is used to indicate the collaboration, the interaction and the communication between two different railway infrastructure models. In fact the combination between two models which represent railway network at different levels of detail can be implemented to overcome applicability limits relative to models. For example a macroscopic model could be combined with a microscopic model to obtain an integrated model which employs the macroscopic one to compensate for the computational inefficiency of the microscopic model, and the microscopic one to compensate instead for the inaccuracy in results of the macro model. Actually only rare applications of this type can be found in literature. In particular it is worth mentioning the experience of the University of Hannover that in collaboration with *RMCon* and the Austrian Federal Railways *ÖBB*, has realized a bidirectional communication interface between the macroscopic model *NEMO* and the microscopic model *RailSys* (Kettner, Sewczyk and Eickmann, 2003). The main objective that Austrian Railways has achieved thanks to such integration, regards the possibility of verifying the existing microscopic infrastructure against the forecasted amount of railway traffic (returned by a macroscopic simulation). In fact the infrastructure is firstly transferred into the macroscopic model *NEMO*, then this one requests running times and minimum headways for each considered train model from *RailSys*. Therefore using these data and the forecasted traffic demand, *NEMO* estimates traffic flows for each network section

and detects bottlenecks. In this way, it is therefore possible to detect potential bottlenecks in the existing microscopic infrastructure due to traffic increases many years in advance.

In particular the integration between these models consents also the simplification of infrastructure data maintenance avoiding redundant data collections (e.g. for constructing both the macroscopic and the microscopic model of the same network), since the macroscopic model in NEMO, is automatically built using the microscopic network representation implemented in RailSys (Figure 46). This procedure is simply realized converting all microscopic elements such as junctions, stations and shunting yards in nodes of the macroscopic graph, while inter-station tracks are converted in

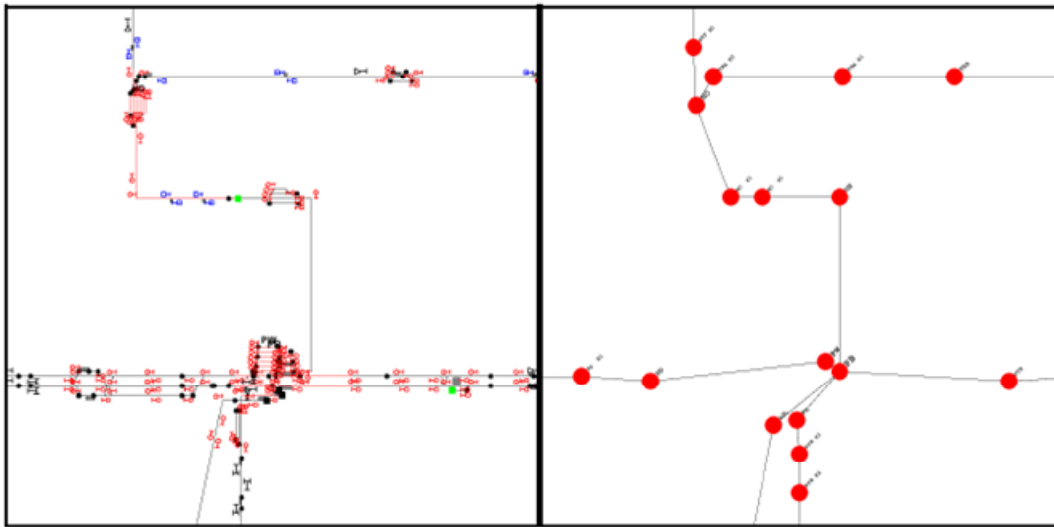


Figure 46. Automatic generation of the macroscopic NEMO network starting from the microscopic RailSys network (Kettner et al. 2003).

links. The communication between the two models has been developed by means of a program-internal interface which automatically implements the exchange of data at high accuracy level (Figure 47). This interface allows the cooperation between models, considering the microscopic model as a server-application acting in the background, while the macroscopic model requests all network infrastructure and operational data and acts therefore as a client. In summary, beyond the automatic network abstraction, three other main tasks are performed thanks to the interaction interface: generation of train models, calculation of running times and determination of minimum headways.

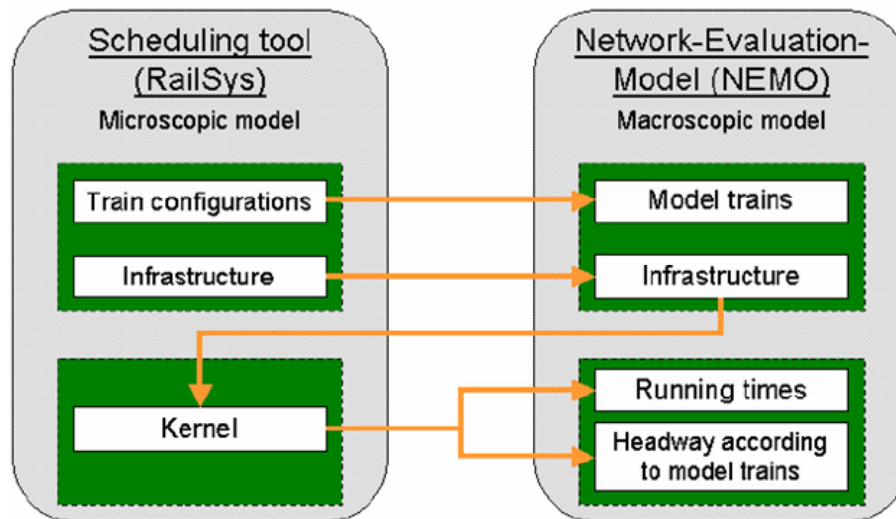


Figure 47. Interaction interface between RailSys (server) and NEMO (client) (Kettner et al. 2003).

In particular for exchanging running times data NEMO must send to the server a request in which the train model (generated according to the different train categories specified within RailSys model) and the corresponding train path have to be specified. Then RailSys receives this message and calculates for that kind of train model and the requested paths four alternative running times according to different train behaviours: a) neither stop at the beginning nor at the end station, b) stop only at the beginning station, c) stop only at the end station, d) stop at both stations.

For determining minimum train headways, NEMO identifies those contiguous route sections where trains can neither pass nor overtake each other. This section are in fact sent to the server, together with information about train models and their path, and subsequently RailSys calculates the minimum headway for each one of the following combinations: a) train 1 and 2 run in the same direction $A \rightarrow B$, b) train 1 and 2 run in the same direction $B \rightarrow A$, c) the trains run in opposite directions (1 in $A \rightarrow B$, and 2 in $B \rightarrow A$), d) the trains run in opposite directions (1 in $B \rightarrow A$, and 2 in $A \rightarrow B$).

As can be seen, hybrid infrastructure models not only allows to overcome applicability limits of the single models but gives the possibility of obtaining enhancements in the estimation of network performances and potentialities.

3.3. Synchronous simulation models

The term “synchronous” stays to indicate that the simulation of a certain model is realized following a time-driven approach where all components are updated at the same time according to events whose occurrence is in the same order as in reality. Therefore the total simulation period is subdivided in a certain number of elementary time intervals (which logically depends on how large such time intervals are), so-called time-steps, and for each one of these time step, the characteristics and performances of each component of the system are calculated. Practically, for the specific case of railway systems both rail vehicle characteristics (e.g. speeds and positions) and positions of signalling equipments (e.g. signal aspects, switches positions) are updated at each simulation step in order to know for each time instant the status of the overall system and therefore identify influences amongst the different train running on the network. There is no “roll-back” in the chronological progression (though in theory it is possible), and the system has to react immediately (or with a certain delay) to every situation. Therefore as said before, in a synchronous simulation, all events happen in the same order as in reality. Here the kernel of the simulation handles the simulated trains and updates their status, according to the results of several subsystems and with the progress of time. In particular, the status of time-driven system components such as rail vehicles are calculated at each time step considering also their status at the preceding instant, therefore train status is updated making use of the chained equations already presented at (1b). While the status of event-driven components like control components (signals, switches, interlocking equipments) is updated in correspondence to the occurrence of an event. Therefore the aspect of a line-side signal will change for example when a train occupies a certain block section, or when an operator forces its aspect to allow special train movements on shunting yards, and so on. In particular to each one of such events a specific time mark can be assigned in order to fix its occurrence time. Therefore each event can be saved in a chronological queue, and at each simulation step the event having a time mark equal or lower than the current simulation time, is simulated. For this kind of system components, their status is updated therefore by using equations of the type already reported at (2).

Specifically, many synchronous models have been developed for the microscopic simulation of railway network operations. In fact the aforementioned microscopic model *OpenTrack* and *RailSys* belong to this class. Moreover other synchronous

simulation models can be mentioned such as: *VISION* and *RailPlan* developed throughout the UK, *FALKO* and *TRANSIT* distributed by Siemens, as well as *RAILSIM* distributed by Berkley Simulation Software in the USA.

3.4. Asynchronous simulation models

As already anticipated within the introducing section, another simulation method is the so-called “asynchronous” approach. This term mainly stays to indicate that not all components of the system are updated at the same time, but the order for calculating their status is given by established criteria. In particular, for the specific case of microscopic railway simulation, this kind of models requires as input data an initial timetable, exact running times and all information about the kind of signalling system implemented on the considered network. Here not all trains are simulated at the same time (as in the synchronous simulation), but the order through which each train is calculated must follow a certain criterion, often relative to the category to which the train belongs to. In fact the simulation starts by modelling trains with high priority, and in general practice long distance passenger trains have the highest priority while urban freight train has the lowest. Therefore trains with higher priority are only influenced by disturbances which happens to themselves like longer dwell times or technical failures, since they are simulated before lower priority trains, and for this reason in an asynchronous simulation they are not influenced by these ones. On that account trains with high priority experience only rather small consecutive delays in an asynchronous simulation. Trains with low priority instead are simulated after high priority trains and in particular fit into time intervals that are left after the calculation of the high priority trains. As can be clearly seen, asynchronous simulation is more static than synchronous simulation and it is especially suitable for conflict-free timetable construction, since it reproduces the process of timetable design.

Moreover if an asynchronous simulation model is employed to analyze railway operations, trains with high priority are always preferred and displace trains with lower priority. Because of the strictly hierarchical structure of the asynchronous simulation, trains with low priority may receive therefore more delays than they would in reality.

Then, the fact that at every simulation step the state of every individual element (e.g. train, signals) of the simulated reality is known, enables the synchronous simulation to be more flexible than asynchronous simulation models. Here, the status of processed

trains is in fact not altered after they are inserted into the simulation timetable. That means a lower priority train would only be assigned a higher priority in order to avoid a larger delay if appropriate dispatching rules were implemented. A higher priority is however often required if train delays are higher than a certain threshold, to ensure that the dispatcher takes actions to bring a delayed train back to schedule as soon as possible. Some asynchronous models have been recently developed to satisfy in particular the need of conflict-free timetable design. Among examples of such model, it is possible to mention *BABSI* (Gröger, Franke, 2002) and *STRESI* (Shultze, 1985), both developed by the RWTH of Aachen (Germany).

3.5. Deterministic and Stochastic simulation models

As already anticipated within paragraph 1.1, simulation models can be classified in deterministic and stochastic according to the assumption made for time instants into which considered events do occur (Figure 48). In fact if the occurring time τ_i associated to the event e_i is considered as a constant variable then the simulation will be classified as deterministic. If τ_i is assumed instead as a random variable having a certain probability density function, the simulation will be classified as stochastic.

Therefore, referring to the case of the simulation of railway operations, it is possible to define a deterministic model, if events like train arrivals, departures or running times are constant variables equal to that values established by the timetable. These kinds of models are mainly used for the preliminary design of timetables to check scheduled conflicts due to the overlap of train blocking times. In addition deterministic simulation can be performed also to verify if system components meets the requirements needed for implementing scheduled operations.

When instead arrival and departure times, dwell times, or running times are considered as random variables, it is possible to talk about a stochastic railway simulation model. This kind of models are principally applied for detailed design of timetables, for estimating their robustness against operation disturbances, for testing network stability, and modelling individual train movements controlled automatically by signals, coils and on-board processors. Usually, stochastic disturbances are included within microscopic railway simulation models considering train dwell times at station as a random variable. In particular some studies have been conducted in literature consisting in measuring real station dwell times of both urban (metro) and long distance trains in order to correctly

model such variables during simulation, therefore increasing the reliability of results returned by timetable robustness investigations and/or network stability analyses. Specifically such studies show for long distance passenger trains that their free dwell times of late arriving fit well with a Weibull probability function (Yuan, 2006), while for metro trains their station stop times fit best with a log-normal distribution (Martinez *et al.* 2007).

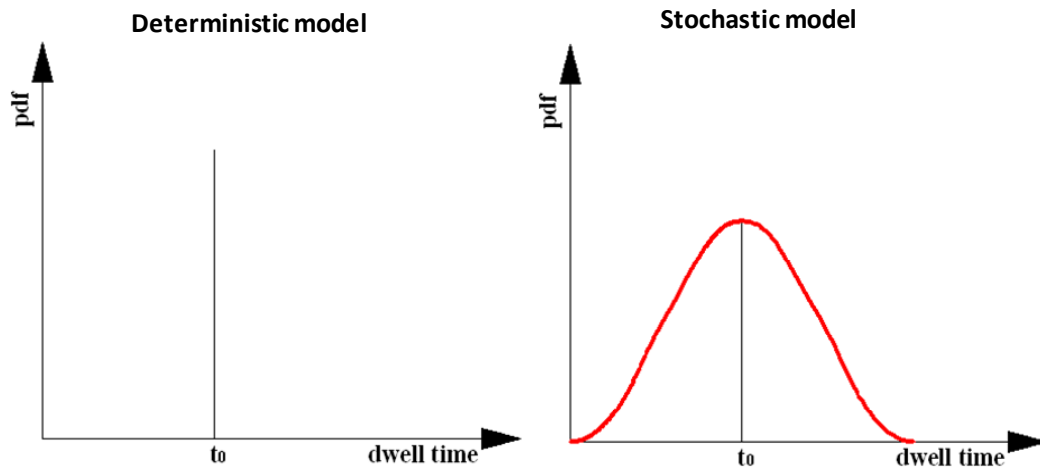


Figure 48. Assumptions made for a deterministic and a stochastic simulation model.

3.6. Applications of railway simulation models for supporting design activities.

Design and planning activities, strongly need to be supported by simulation models of railway operations, to accurately evaluate effects induced on system performances by a certain intervention solution. In particular such models can be employed within different ways in order to meet the requirements of the specific case study under investigation. For example in literature several applications can be observed which use simulation models of railway networks to realize simple “what-if” evaluations (therefore implementing a simple simulation of the considered scenario) addressed to verify both infrastructure layouts and timetables under scheduled operations, analysis of timetable robustness and/or network stability (considering stochastic disturbances to trains during simulation), design of conflict-free timetables, or the verification of capacity values obtainable by implementing an optimized layout of signalling system. Moreover, many applications have been also carried out for the designing of robust timetables, or to identify optimal train rescheduling patterns to mitigate effects of train

disturbances within real-time operations. However, since the implementation of these latter kind of applications require for the first case a large number of model simulations, and for the second case reduced computing times (because of the real-time application), it has been necessary to rely on “simplified” microscopic models (e.g. queuing models, fixed-speed models) that allow to obtain low computing times but with less accurate evaluations of system performances.

In the following sections a general overview will be supplied on the mentioned applications observed in literature, and a brief description of “simplified” models developed to obtain more efficient computing times will be realized.

3.6.1. Verifying the stability of timetable and network

Middelkoop and Bouwman (2002), had applied the macroscopic infrastructure model SIMONE (previously introduced), to support the verification and testing phases of both timetable and infrastructure of a certain section belonging to the Dutch railway line. In particular it is possible to consider a timetable as stable (or robust), when it is designed in such a way that small random disturbances to operations can be absorbed by recovery or buffer times of the timetable, preventing the propagation of delays on the network. In the same way, it is possible to call a network as stable when its layout has been designed in such a way that slight disturbances experienced by a train during operations, do not propagate back to following trains. Therefore the application carried out, consisted in simulating the considered railway section under disturbed operational conditions in order to verify how robust and stable the configurations of both scheduled plan and network infrastructure were.

Specifically stability was measured through measuring a set of performance indicator like: the number of delayed trains, punctuality, used recovery-time (i.e. slack in running and dwell times), the ratio of introduced disturbances related to resulting delays, broken connections, delay absorption.

A first application was realized investigating how the designed railway system was stable with respect to disturbances occurring to connecting times between two connecting regional trains at the Dutch railway station of Weesp (Figure 49), supposing that a connection is broken when the waiting time of a train for connecting to the other becomes higher than 3 minutes. In this experiment, disturbances are added supposing several disturbed scenarios. Results showed that the designed network was partially

sensitive for the variation of the waiting time experienced by a train to connect with the other one at the Weesp station, since in case of increased waiting time an appreciable decrease in train punctuality is estimated.

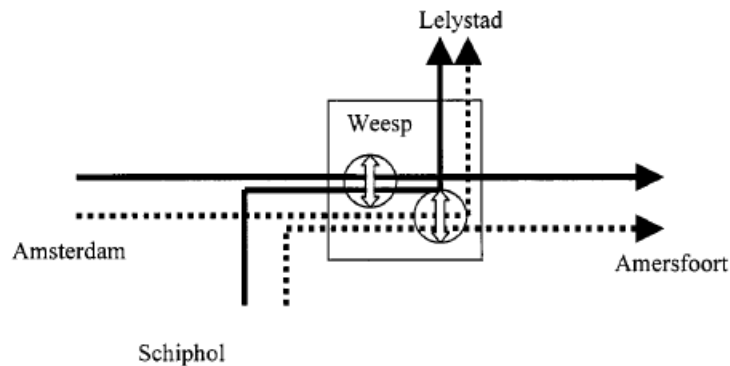


Figure 49. Train connection in Weesp station (Middelkoop, Bouwman, 2002).

Another application was then referred to examine the stability of the network when 6 extra freight train paths are inserted into the timetable. Under this design hypothesis, system stability was investigated evaluating network performances for several disturbed scenarios of the considered operations, introducing different kinds of noise disturbances to all train running times and dwell times. Results showed that adding the 6 train paths, an increase of 10% of the average train delay is observed in the network (Figure 50), with respect to the original timetable. In particular, punctuality was decreased, mainly because of smaller headway times when further 6 trains are considered.

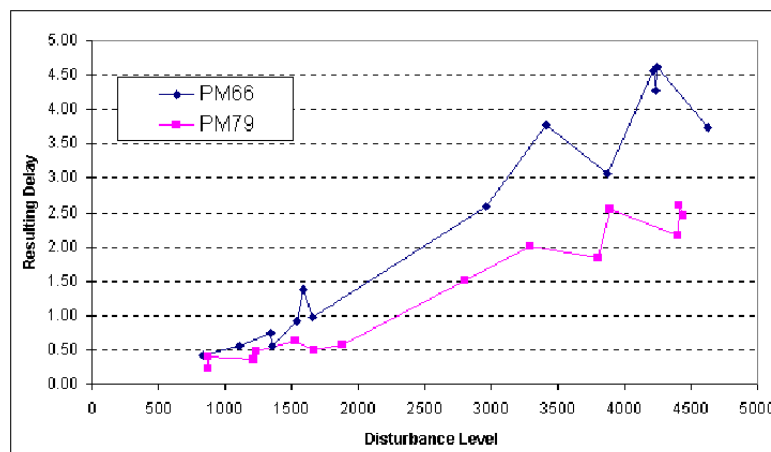


Figure 50. Resulting delay for non intervention (pink line) and intervention (blue line) scenarios for different noise disturbances conditions (Middelkoop, Bouwman, 2001).

Moreover a further research was conducted to determine the benefits (in terms of train punctuality) due to the introduction of a switch within a connecting station where in the

current situation two connecting trains are forced to leave from that station with a time difference higher than 2 minutes, since the absence of a switch does not allow earlier train departures. Therefore the presented intervention scenario was simulated to estimate corresponding system performances, within different disturbed conditions (imposing different values of dwell time and running times disturbances) and supposing that the connections between the two considered train lines are broken if one of these trains has to wait for the other coming, more than 3 minutes. Results showed that the introduction of a switch between the rail tracks returned a positive impact in reducing train departure delays (Figure 51a), but a negative effect on the percentage of broken connections (Figure 51b).

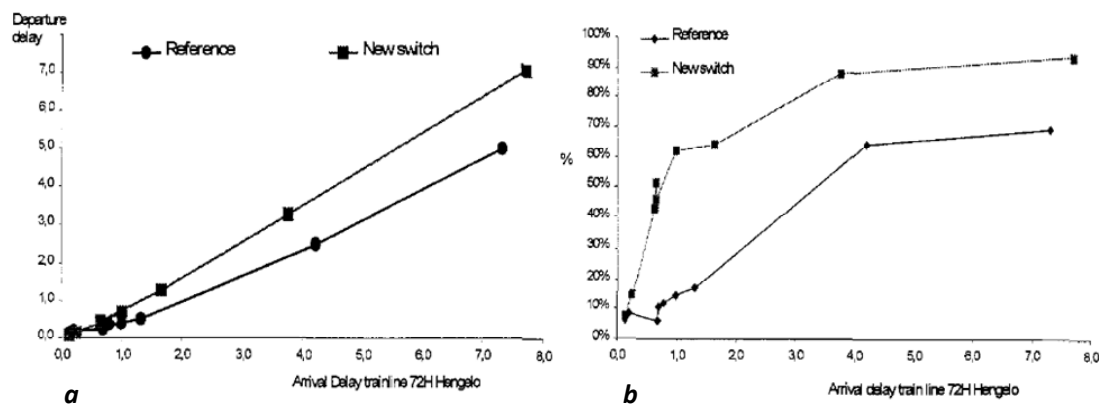


Figure 51. Resulting departure delays (a) and percentage of broken connections (b) for the current and the intervention scenario for different noise disturbances conditions (Middelkoop, Bouwman, 2002).

3.6.2. Verifying system performances for an optimized signalling layout

Gill and Goodman (1992) developed a computer-based optimization technique to design the layout of a multi-aspect equi-block signalling system for mass rapid transit networks which consents to maximize system capacity (i.e. reducing the signal headway and therefore line headway). This method consisted of a first phase in which the value of the undisturbed train braking curve (to come at a standstill from the maximum line speed) was minimized with respect to ATO/ATP speed codes. Then such minimum value was considered to be equal to the block section length of an equi-block signalling system layout, and the signal headway obtainable implementing such optimum configuration was estimated after a microscopic and deterministic simulation of the considered configuration. To keep low computing times of the optimization process, the analytical form of the objective function (i.e. the undisturbed train braking distance under multi-aspect signalling rule), was considered. As already shown within the previous chapter,

the undisturbed braking curve of a train controlled by a multi-aspect signalling layout has the shape illustrated in Figure 52. In particular it is possible to express the braking distance $f(v_i, v_{i+1})$ of the reported curve necessary to decrease speed from the ATO speed v_i to the ATP speed v_{i+1} , in function of these speed codes, with the following equation:

$$f(v_i, v_{i+1}) = t_r v_i + \frac{b_s}{2J} (v_{i+1} + v_i) + \frac{1}{2b_s} (v_i^2 - v_{i+1}^2) \quad (24)$$

where t_r is the delay of the communication system relative to signalling equipments, b_s is the service train braking rate, while J represents train jerk rate (i.e. the variation of deceleration rate during braking).

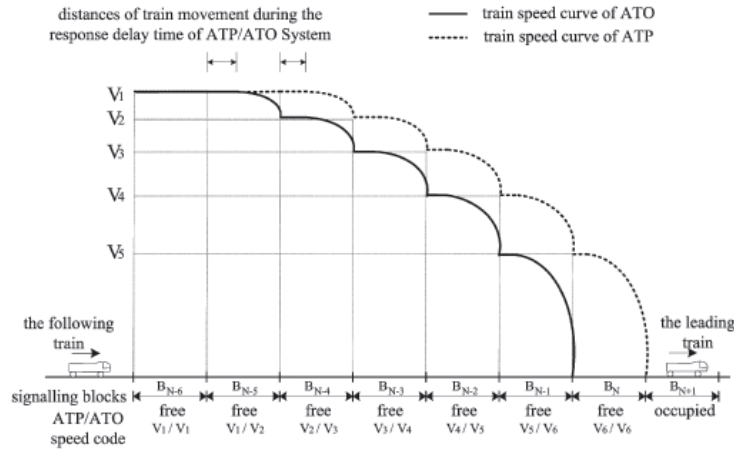


Figure 52. Train braking curve in a six-aspect equi-block layout and relationships with ATO/ATP speed codes (Ke et al. 2009).

Therefore, for instance in the case of n -aspect signalling layouts the objective function which must be minimized, is given by (25) and represents the sum of the differences between braking distances relative to contiguous block sections. In fact, the pattern of ATP/ATO speeds ($v_1, v_2, v_3, \dots, v_{n-2}$) which minimizes function F returns a n -aspect equi-block signalling layout which allows a train to safely brake from maximum line speed until coming at a standstill.

$$F(v_1, v_2, v_3, \dots, v_{n-2}) = (f(v_m, v_1) - f(v_1, v_2)) + \sum_{i=1}^{n-4} (f(v_i, v_{i+1}) - f(v_{i+1}, v_{i-2})) + (f(v_{n-3}, v_{n-2}) - f(v_{n-2}, 0)) \quad (25)$$

where v_m is the maximum line speed, while v_k for $k = 1 \dots n-2$ are the ATP/ATO speed codes. In particular once the optimal speed codes pattern has been obtained, the equi-block section length is calculated for one of the considered block section using equation (24).

Moreover it is necessary to say that station areas are the most critical sections that influence the minimum line headway of a railway line (and therefore its capacity), since here trains have to stop occupying related block sections for a time that is higher than open line tracks. Therefore, to reduce line headway is enough to implement the described equi-block layout only within station areas, while on the open line such block lengths can be enlarged to reduce the number of track circuits and therefore installation costs. Hence, according to the convention used by Chang and Du (1998), it is possible to define a *constrained* section, a *critical* section, and a *stretchable* section (Figure 53).

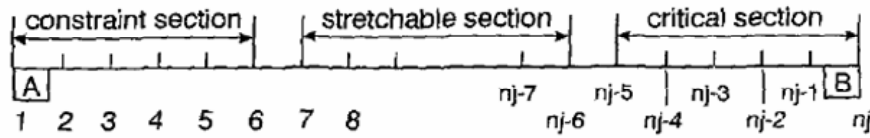


Figure 53. A 4-aspect equi-block section layout for a metro line (Chang and Du, 1998).

The first two kinds of sections which are located within station areas must have the designed block section length to guarantee an increase in line capacity, while the stretchable section characterizes the open line track, therefore its block sections can be enlarged to reduce the number of track circuits and line side signals.

Gill and Goodman solved the optimization problem using a gradient-search technique. Once the solution was identified and therefore the lengths of constrained and critical sections were found, a post-optimization procedure was carried out to enlarge block section lengths of the stretchable section. This was realized satisfying some constraints such as the congruence amongst starting and ending joints of contiguous block sections. Then, obtained block section layout (containing all inter-station sections: constrained, critical and stretchable), was verified with respect to the ATP emergency braking, checking that train braking distance within emergency conditions (therefore using a braking deceleration rate equal to $b_e \approx 1.2 b_s$) for each block section is lower than the corresponding block section length.

Chang and Du (1998), instead brought an improvement to the method developed by Gill and Goodman, in fact they solved the whole block section design problem, defining for each one of the three type of section (constrained, stretchable and critical), a different objective function as reported in (26), (27), and (28).

$$f_{constrained} = \frac{1}{\sum_{i=1}^{n+1} (f(v_i, v_{i+1}) - (x_{i+1} - x_i))} \quad (26)$$

$$f_{critical} = \frac{1}{\sum_{i=nj-5}^{nj} (f(v_i, v_{i+1}) - (x_{i+1} - x_i))} \quad (27)$$

$$f_{stretchable} = \frac{1}{1.1^{N_{new}} \cdot fit_a \cdot fit_b} \quad (28)$$

where n represents the number of constrained sections, n_j the number of critical sections, $f(v_i, v_{i+1})$ is the service braking distance for each block section, x_{i+1} and x_i are respectively the position of end and begin joint of the i^{th} block section. N_{new} constitutes the number of block sections belonging to the stretchable section, fit_a is a function assuring that the enlargement of block sections does not increase the minimum line headway, while fit_b guarantees that the length of each block section is higher than the corresponding ATP emergency braking distance.

As is clear, here the design variables are block section joint positions for each kind of inter-station section. Moreover, with this approach no post-optimization procedures to enlarge block section lengths of stretchable section, neither to verify block lengths against emergency braking distance, are necessary. However, a genetic algorithm was used to find optimal solutions for each one of the reported objective functions (Chang and Du, 1998).

A further improvement to such method was then brought by the same authors Chang and Du in 1999, which combined in one single composite objective function the three different objective functions reported in (26), (27), (28). In particular this overall objective function is expressed as in (29):

$$Obj(x) = \sum_{i=1}^{n+1} |f_i - x_i| + \sum_{i=d_j-(n+1)}^{d_j} |f_i - x_i| + Penalty_a + Penalty_b + Penalty_c \quad (29)$$

where f_i is the service braking distance of the i^{th} block section (as given by (24)), x_i is the length of the i^{th} block section, $Penalty_a$ is related to the minimum headway and the number of blocks in the stretchable section, $Penalty_b$ is related to the emergency braking test in the stretchable section, while $Penalty_c$ is related to the emergency braking test in both the critical and the constraint sections.

The optimal pattern of block section lengths x_i which minimizes the objective function $Obj(x)$, was found through using a differential evolution algorithm for improving computing times of the optimization process. In fact the use of this algorithm and the composite objective function reported in (29) slightly reduce computational times to find the optimal solution also with respect to the GA-based design approach.

Ke et al. (2009), formulated instead another method to solve another kind of problem consisting in the design of a multi-aspect equi-block layout which minimizes train energy consumption during operation. In particular, all the length of the inter-station track is initially subdivided in equi-block section whose lengths and ATO/ATP speed codes v_d are given, minimizing the objective function (25). Once that the optimal equi-block layout has been obtained, a further optimization problem is solved to identify the pattern of average speeds v_k that trains can adopt on the k^{th} block section to minimize energy consumption respecting the constraint for which $v_k \leq v_d$, for each section. The energy consumed E_k by a train running on a block section k is given by the integral on time T of the instantaneous power (returned as the product between the average running speed v_k and the average active force F_k used to move the train on section k):

$$E_k = \int_0^T F_k \cdot v_k \cdot dt \quad (29)$$

Therefore the objective function to solve the described problem is given by:

$$\min \sum_{k=1}^n E_k(v_k); \quad \text{for } v_k \leq v_d \quad (30)$$

Where n is the number of block sections composing the inter-station track and v_d are the corresponding ATO/ATP speed codes. The pattern of optimal block section average

speeds v_k which minimizes train energy consumption, is found by using a Max-Min ant system of Ant Colony Optimization algorithm.

Anyway, once optimization solutions have been obtained and therefore optimal values (of block section lengths, ATO/ATP speed codes and/or average block speeds) have been determined, the corresponding network performances can be only evaluated by using a microscopic simulation of railway operations. In fact, using as input data of a synchronous simulation model, the optimal values of equi-block section lengths and speed codes, the minimum line headway (and therefore line capacity) is evaluated through simulating train runs and applying the blocking time theory (Gill and Goodman, 1992, Chang and Du, 1998,1999). The same thing was realized by Ke et al. (2009) who employed a synchronous microscopic simulation model to estimate train energy consumption values in correspondence to block section lengths and average speeds obtained as solutions of the optimization problem (30).

3.6.3. Designing of robust timetable

As already said, a timetable can be defined as robust, when its supplement times (recovery to train running times and buffer times between consecutive train departures) are designed in such a way that small delays during operation are absorbed in a certain measure, preventing their propagation in the network. Therefore as seen in the previous sections, to test the robustness of a given timetable it is necessary to simulate several disturbed scenarios of railway operations scheduled according to that timetable. In particular Kroon et al. 2007, developed a method to optimize the allocation of time supplements in a given cyclic timetable through an iterative improving process based on simulating the timetable under stochastic disturbances → evaluating the overall train arrival delay → optimizing the allocation of time supplements through the minimization of train arrival delays.

Specifically this method makes use of a so-called Timetabling Part and a Simulation Part. The *Timetabling Part* models a given timetable as a Periodic Event Scheduling Problem (PESP) as made by Serafini and Ukovich (1989). Therefore, considering that the planned time for an event e is given by parameter v_e , it is possible to represent planned process times between to events e, e' as:

$$M_{e,e'} + s_{e,e'} = v_{e'} - v_e; \quad \text{for all } (e,e') \quad (31)$$

Where $M_{e,e'}$ is the minimum process time between event e and e' (e.g. the minimum running time between departure at station A and arrival at station B), $s_{e,e'}$ represents supplement time for the considered process (e,e') , while $v_{e'}$ and v_e are respectively the end and the beginning time of process (e,e') (e.g. the arrival time at station B and the departure time from station A). Therefore a cyclic timetable is represented by equations of the type illustrated in (31), coupled with constraints such as:

$$0 \leq v_{e'} - v_e \leq T \quad (32)$$

$$\sum_{(e,e')} s_{e,e'} \leq Sq \quad (33)$$

where in fact constraint (32) assures that the planned process (e,e') is contained within the cycle duration T (usually considered as 1 hour), while constraint (33) guarantees that the sum of time supplements over all planned operations (e,e') must be lower than a certain amount Sq , in order to prevent a reduction in capacity utilization of the railway line.

The *Simulation Part*, instead is used to realize R replications of disturbed train operations, (adopting for each replication r a certain disturbance pattern) as scheduled by the Timetabling part, in order to evaluate how robust the designed timetable is with respect to these stochastic disturbances. In particular to reduce computational time for evaluating the timetable, the simulation model here employed is not an infrastructure model but a macroscopic model based on activity graphs. Since it is a macroscopic model only events related to stations or network junctions will be considered (e.g. train arrival/departure at/from station or important junction nodes) neglecting therefore events related to lower level elements (e.g. train arrivals at block section joints). A sample of this kind of models is reported in Figure 54, where train operations are modelled as processes (e,e') whose starting and ending times are respectively given by parameters w_e and $w_{e'}$.

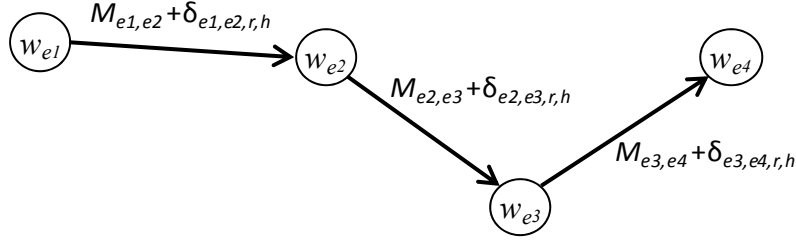


Figure 54. Stochastic activity based macroscopic simulation model as used in the Simulation Part of the Stochastic Optimization method developed by Kroon et al. 2007

Specifically such time instants are random variables since they are constituted by a deterministic value $M_{e,e'}$ which is the minimum processing time for process (e,e') (e.g. minimum running time on a certain inter-station track) and a random value representing a stochastic disturbance which also depends on the replication r and the timetable cycle h in which operation is simulated. Such model can be therefore represented into a mathematical form by equations of the type:

$$M_{e,e'} + \delta_{e,e',r,h} = w_{e',r,h} - w_{e,r,h} \quad \text{for all } (e,e') \quad (34)$$

Hence, for each one of the R simulation replications, a certain configuration of the disturbance pattern $\delta_{e,r,h}$ is drawn from a negative exponential distribution, and the corresponding train arrival delay $\Delta_{e,r,h} = w_{e,r,h} - v_e$, at stations for replication r and cycle h is evaluated. To improve the robustness of the simulated timetable, it is therefore necessary to minimize the weighted total (or average) train arrival delay:

$$\min \Delta = \sum_{e \in Ea} \sum_{r=1}^R \sum_{h=1}^H c_e \cdot \Delta_{e,r,h} \quad (35)$$

where Ea is the total number of train arrival events at line stations, R is the number of total replications, H is the number of timetable cycles, while c_e is a weight for different delays since for passenger a train delay at a station could be more harmful than a delay at another location.

Once this optimization problem has been solved, and therefore a new pattern of supplement times $s_{e,e'}$ has been determined, this improved timetable will be simulated again to test its robustness, and another optimization is performed. Then, the described improving process continues iteratively, until no more timetable improvements are obtainable.

3.6.4. Real-time rescheduling of train operations.

When during actual operations, train conflicts due to stochastic disturbances (e.g. element failures, longer station dwell times, etc.) are detected, it is often necessary to reschedule delayed train runs, in order to mitigate impacts of delays on other trains. In these cases, dispatchers have the responsibility of identifying in advance all possible train conflicts and determine an effective rescheduling pattern through which operations can return as soon as possible to ordinary conditions. Given the complexity of this task, the support of simulation model for the prediction and the resolution of conflicts, is strongly needed. In fact, a high-quality rescheduling action should be performed carrying out the following main functions:

- Start with a conflict-free timetable
- Get information from the operation
- Detection of conflicts
- Automatic resolution of conflicts
- Generation of a new conflict-free dispatching timetable
- Provision of data for traffic control

Moreover, rescheduling actions could be addressed to reach one or more objectives regarding the mitigation of disturbances effects on the network. In particular such objectives can be for example: the minimization of the overall train arrival delay, the minimization of a weighted delay (based on a priority for each train), or returning as soon as possible to the original timetable.

However, tools for supporting real-time rescheduling activities are composed by two different modules:

- *Conflict detection module*, which detects (or predicts) through simulation of train runs, all the conflicts amongst trains on the network for a certain time period.
- *Conflict resolution module*, which solves all detected conflicts giving as output a new timetable with fewer or no conflicts, according to the modification rules adopted.

As already said, the conflict detection module is usually constituted by a microscopic simulation model, since the accurate calculation of train trajectories can be made only when a detailed representation of network infrastructure (gradients, radii, track speed limits) and signalling equipments is available. Then train time-space trajectories are employed to estimate blocking times, in order to find allocation conflicts due to overlaps of train blocking times. Moreover, in addition to allocation conflicts also connectional conflicts due to delays of connecting trains, must be detected.

The conflict resolution module, instead must solve all conflicts detected by the previous module in order to return a new conflict-free timetable. Different approaches can be employed to achieve this purpose. In particular the three main used approaches are: asynchronous simulation, synchronous simulation or optimization methods.

Asynchronous simulation

As seen before, asynchronous simulation models are mainly used to design conflict-free timetables, preventing overlaps among train blocking times. For this kind of conflict resolution approach, it is necessary the specification of train priorities, since here train runs are not simulated all together following the sequence of events as in reality, but trains are simulated following the order given by their own priority level. Usually, long distance passenger trains have the highest priority, while local freight trains have the lowest. Hence, if the conflict detection module identifies a conflict between two trains, the asynchronous module proceeds to solve all allocation overlaps between such trains as well as knock-on conflicts with other trains. The scheduling procedure realized by such kind of model consists in simulating trains with first-rank priority and scheduling their paths in a conflict-free way, then all trains having second-rank priority are simulated and inserted into time lags between higher priority train paths, avoiding conflicts. Then, the same procedure is carried out to allocate train paths with lower priority, until a new conflict-free timetable is returned.

Moreover some asynchronous models, such as the tool ASDIS (Jacobs 2003, 2004), consents to the dispatcher to identify which operative strategy to use to solve train conflicts. In particular, such options are: a) use of alternative routes, b) extension of a scheduled stop, c) relocation of passing stops, d) additional stops for operational requirements, e) extension of running times. However independently of the chosen

alternative, the result of the asynchronous simulation model is a conflict-free dispatching timetable with no block occupation conflicts.

Synchronous simulation

Synchronous models recreate operating processes simulating events in the same chronological sequence as in reality, using the traditional time-step simulation. Anyway, since synchronous models simulate all trains present on the network at the same time, they preclude any automated generation of non-conflicting running schedules. Actually, it is only possible to predict in advance train conflicts. Therefore the use of these models during dispatching activities, basically regards the prediction of conflicts and the simulation of different operational strategies according a “what if” approach, to identify and put directly into operation, the one which better mitigates disruption effects on the network.

Optimization methods

Recently, some methods have been developed to build conflict-free timetables which satisfy certain operational objectives such as for instance the minimization of the overall train arrival delay, the minimization of a weighted delay, or the minimization of the time to return within ordinary conditions. Therefore, this approach needs a certain objective function which must be minimized, respecting some operational restrictions (e.g. connections between two trains at a station, train movements) which are considered in the model as mathematical constraints. Since this task has a high complexity, the number of constraints must not be large and the simulation model has to be efficient from the computational point of view, in order to guarantee a calculating time which is acceptable also for managing real-time operations. Usually, branch and bound search techniques are used to identify a first solution and then such solution is improved employing other methods. For example, Wegele and Schnieder (2004) used a combination of a branch and bound with a genetic algorithm which was used to progressively improve initial solutions by coupling several solutions (cross-over).

D’Ariano et al. (2006, 2007, 2007) for example developed an optimization approach based on modelling train operations at a microscopic level using the so-called alternative graphs. This kind of model represents train operations as activities (or processes), and since it is microscopic, all train interactions with both main infrastructure components (e.g. arrival at stations and junctions) and signalling

equipments (e.g. train arrival at block section joints) are considered. Figure 55 shows a typical representation of a small railway network according to an alternative graph formulation. Here the length of a link can depict for instance the running time of a train on a block section, while a node can represent the departure time of a train from a station or the arrival time at a block section joint.

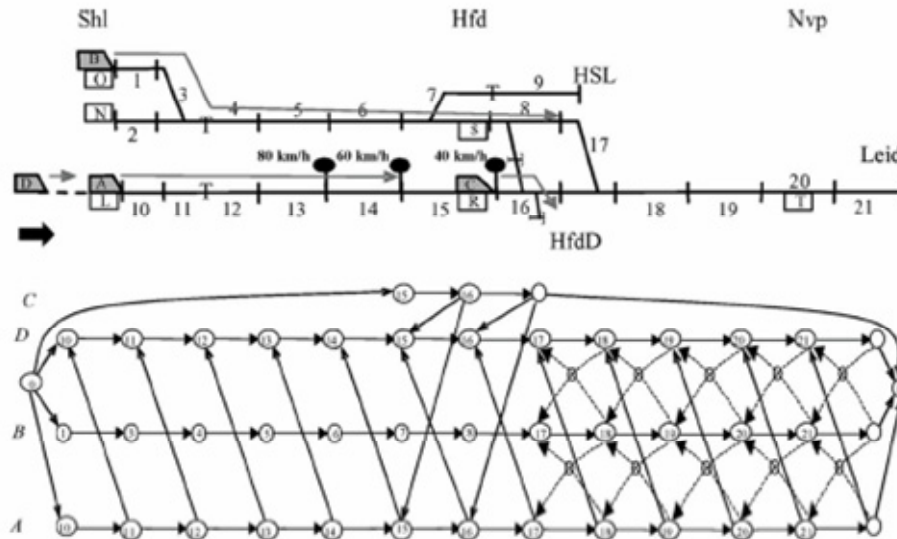


Figure 55. Representation of a small network as an alternative graph (D'Ariano et al. 2007).

Furthermore, these models are considered as *fixed-speed* models, since they assume that train travel times on a block section are deterministic parameters, whose values correspond to the undisturbed running times (i.e. as scheduled by timetable). Therefore such models are lacking in results accuracy when disturbances to train movements (e.g. an unforeseen reduction of speed due to a yellow signal aspect) do occur, since in these cases, they are not able to reproduce train transient phases (e.g. deceleration and acceleration) and estimate precise running times and therefore train blocking times. However, in the work presented by D'Ariano et al., a considerable effort has been done to try to solve this aspect, including in the conflict resolution model an iterative train speed updating procedure, which is addressed to adjust train speed profiles when conflicts due to headway or route arise. In summary the architecture of this model is depicted in Figure 56. Practically, after that field data (e.g. positions and speeds of trains) have been loaded as inputs into the dispatching system, the fixed-speed conflict detection and resolution first identifies potential headway or route conflicts. Then it solves such conflicts searching for a conflict-free train configuration, considering all the train speed profiles fixed as scheduled, and aiming at the minimization of consecutive

train delays in the network. The branch-and-bound algorithm is used to solve this optimization problem. In particular at each iteration of the optimization process the conflict resolution module is employed to determine a solution using a train fixed-speed profile. Then a feasibility check is realized to verify if all trains have an acceptable speed profile and above all respect minimum safe distances. If the solution is acceptable, the iterative rescheduling process stops and a final solution is returned to the dispatcher. Otherwise, a speed updating procedure is activated to adjust train speed profiles and a new iteration begins to find a new conflict-free timetable.

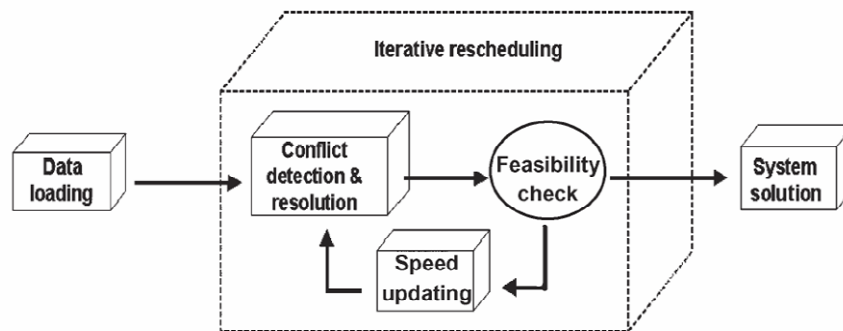


Figure 56. Architecture of the dispatching system (D'Ariano et al. 2007).

A similar approach was adopted by Mazzarello and Ottaviani (2005, 2007) who also developed a conflict detection and resolution module (CDR) based on alternative graphs. In particular, given an initial timetable, a set of constraints (e.g. minimum speed constraints, stop and departure constraints, connection constraints, out of order constraints, etc.) and speed/position coordinates of each train, the CDR automatically detects conflicts and creates a schedule of earliest/latest possible arrival times, departure times and speed for the trains at a certain number of key points. Then the CDR solves all detected conflicts and return an optimized conflict-free timetable by means of the alternative graph (with fixed-speed profiles of trains) using a heuristic algorithm. Then another component called Speed Profile Generator (SPG) is used for the computation of an optimal speed profile for the trains, minimizing a cost function which involves punctuality, deceleration and acceleration.

Chapter 4. Development of a microscopic infrastructure model for simulating railway operations.

4.1. Introduction

The previous chapter has shown the different kind of models for simulating railway operations, according to the level of detail through which the network is represented (macroscopic, mesoscopic, microscopic), the type of analytical approach (deterministic, stochastic) and the technique to process network events (synchronous, asynchronous). Then, it has been underlined the importance assumed by railway simulation models in supporting different planning and design activities for the evaluation of the effects induced by a certain intervention solution on system performances. In particular several applications observed in literature have been presented for both “off-line” design tasks, addressed to intervene on infrastructure and/or operational strategies (e.g. timetable), and “on-line” activities, aiming at the management of real-time train operations to minimize impacts of stochastic disturbances on the network. Moreover, such description has highlighted the importance of accuracy in model results, for correctly estimating system performances within a certain scenario, as well as of computational efficiency, to allow acceptable computing times during real-time rescheduling or probabilistic analyses like robust timetabling and network stability tests. Although a more thorough estimation of network performances is achievable only adopting microscopic infrastructure models, these ones are usually used in practice for evaluating only a limited set of intervention scenarios (for example according to a “what-if” design approach), while it is preferred to rely on less accurate models such as higher abstraction level (macroscopic, mesoscopic) or “fixed-speed” microscopic models, when investigations requiring a large amount of model estimations (e.g. black-box optimization, probabilistic analyses) must be performed. This obviously implies certain approximations in simulation results, which can become also unacceptable when congestion levels on the network increase (*Quaglietta et al.* 2011). However, the reason why the use of microscopic infrastructure models are not preferred for this kind of analyses, is mainly due to the large number of input data involved, which makes computing times be not suitable for simulating large-sized network or for studies entailing a consistent amount of model simulations. In addition, it must not be neglected the fact that the closed structure (i.e. the impossibility of modifying inner functions, or

communicating via API with external applications) of commercial microscopic infrastructure models (*OpenTrack*, *RailSys*, *RAILSIM*, etc.), does not allow an automatic interfacing with these kind of mathematical frameworks (e.g. optimization models), compelling therefore users to develop customized and simplified models accomplishing to their own needs, often with an applicability which is limited to the specific case-studies under investigation (i.e. without a general validity).

Insofar, the objective of the work presented in this thesis concerns the development of an innovative microscopic infrastructure model for simulating railway operations, which is structurally open to be involved within the various activities and mathematical analyses regarding design tasks, and guarantees at the same time accuracy in results, and therefore more reliable estimations of network performances with respect to higher level or “fixed-speed” models. To this purpose, a synchronous microscopic railway simulation model has been developed in C++ using an object-oriented programming technique. The open structure of such model makes it be flexible and able to be employed for different kind of analyses since it can be automatically interfaced with external applications (e.g. optimization or probabilistic analysis frameworks), and inner functions can be modified. Moreover, it has a parallel architecture which makes this model be more efficient from a computational point of view, since computing times of simulation are strongly reduced when running on multi-cores computers. Furthermore, the object-oriented concept on which this model is based, consents to depict in detail each component of railway network (e.g. infrastructure, signalling system, rail vehicles, etc.) defining for each one a set of specific attributes and functions, that allow to achieve more precise descriptions of their real behaviour and respective interactions, and hence more accurate evaluations of system performances. In addition, such model has a general applicability, since it is structured in modules (each one representing a certain component of railway network) whose parameters can be initialized directly by the user through external files (e.g. text files), consenting therefore the modelling of any case-study.

In this chapter a description of the architecture and features of the developed model is provided. In particular, each structural module is illustrated in detail and information about the mathematical models on which they are based on, is given. Then, some information is supplied about the parallel architecture of the model, the parallelization technique and the benefits induced on simulation computing times.

The next chapter instead, will describe different applications of the model developed, and will show model potentialities through the achieved results.

4.2. General features of the simulation model

As mentioned in the previous section, the necessity for a microscopic infrastructure model that can support the different types of activities and investigations realized during design tasks, entails a series of fundamental characteristics that the model must have. In particular these characteristics can be listed as follows:

- ✓ *Flexibility*: the model must be able to respond to user's needs, therefore it can be interfaced directly or indirectly (e.g. via API) with external programs and/or mathematical structures and must allow the modification or addition of inner functions.
- ✓ *Accuracy*: the model must be able to reproduce with an appropriate detail, both features and dynamics of railway components in order to accurately describe the behaviour of railway system.
- ✓ *General applicability*: the structure must be designed to let the model be adaptable to any case-study without changing the source code, but simply specifying the input dataset.
- ✓ *Computational efficiency*: computing times of simulation model must be acceptable, in the sense that it must be reasonable the time needed to perform a mathematical analysis which require for example the simulation of large-sized networks or a large amount of model evaluations.

Each one of the aforementioned features has strongly conditioned both choices and the way of proceeding of activities relative to the development of the model. In fact, the achievement of each one of the listed characteristics, implied each time the formulation of problems which not always admitted an immediate solution.

In particular the first point of the list (Flexibility) has mainly influenced the choice of the programming language into which the model has been developed and above all its structure. In fact the model has been designed in C++, since it is a wide spread language which consents an easy communication with the most part of tools for mathematical analysis as well as other external applications. Moreover the model has an open

structure in the sense that inner model functions can be manipulated or further methods can be added without compromising or varying the original source code.

The second feature of the list (Accuracy) has instead conditioned the kind of programming technique used to design the model and the type of event processing technique. In fact the necessity of reproducing in detail the behaviour of each railway component as well as the interactions arising among them, has addressed the programming approach towards an object-oriented concept. Object-oriented programming, allows in fact to define different object classes and declare for each one of them specific class attributes and functions (also called methods). This advantage, has been exploited during implementation to model as an object each component of the railway system (e.g. rail vehicles, signalling equipments, infrastructure, etc.), and specify for each one, the set of their specific parameters (e.g. speed and positions for vehicles, radius and gradient for rail tracks) as well as functions depicting dynamics of that particular component. Moreover a synchronous technique to process network events has been adopted. Therefore the sequence of events is reproduced in the same order into which they occur in reality. This has permitted to consider the dynamic evolution of interactions amongst system components and better estimate also transient train dynamics. Hence the employment of an object-oriented technique and a synchronous approach, have consented to obtain an accurate modelling of both time-driven (e.g. rail vehicles) and event-driven (e.g. control components) components, and their interactions, favouring a precise depiction of inner system dynamics and therefore more accurate estimation of network performances.

The general applicability to which the developed model has to respond, have determined the architecture of the model. In fact the model is composed of four main modules, each one of them represents a certain component of railway system (i.e. infrastructure, rolling stock, signalling system, timetable). This modular structure consents in fact to adapt the model to any case-study, since the user can directly specify for each module the corresponding input dataset relative to the case-study under investigation, by means of external files (e.g. database organized in text files). This feature therefore makes the model be valid in general avoiding the modification of the source code which instead is needed for a “case-sensitive” model.

Computational efficiency, is one of the main issue which has been considered during the development of the microscopic model. In fact it is known that, given the large amount of input data considered by microscopic infrastructure models, they are inefficient for simulating large-sized networks or for being used within analyses requiring a wide number of model evaluations. However, modern multi-core computers and multi-threading programming, have provided a strong support to solve this problem. In fact many solutions have been studied to try to reduce model computing times, but the one which consented to increase computational efficiency keeping at the same time the results accuracy of microscopic infrastructure models, was the implementation of a parallel architecture. This task has been hard to carry out and has required a certain effort for studying the problem and concretely implementing the solution. In fact the implementation of a parallel architecture has entailed the need of modifying the structure of many functions and class methods, rewriting them in order to be correctly interfaced with the parallelization paradigm OPEN MP. Hence, this technique has allowed to obtain computing advantages which increase with the increasing of network dimension and the number of cores with which the computer is equipped.

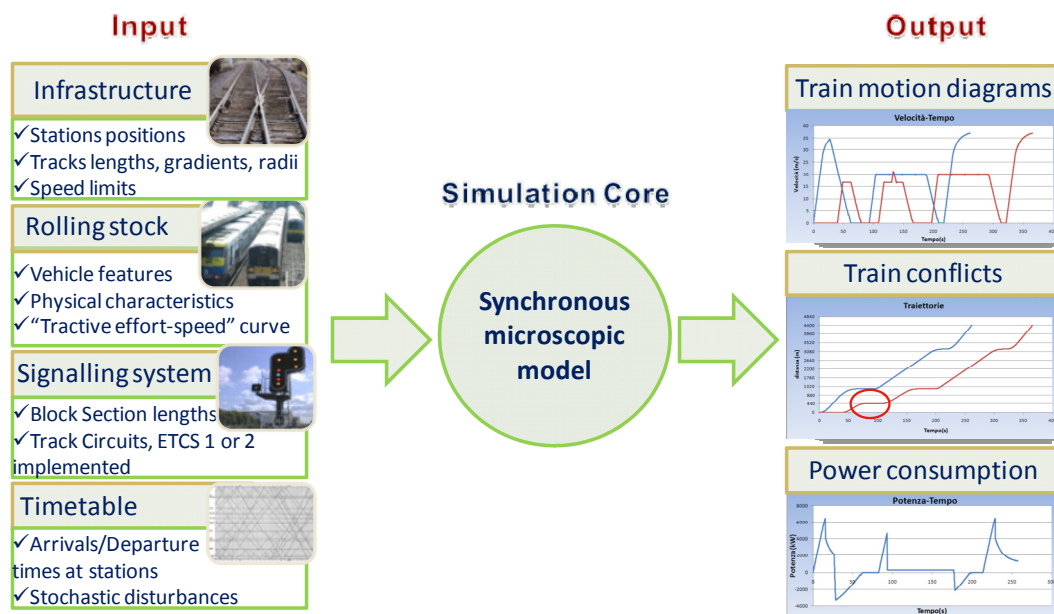


Figure 57. Architecture of the synchronous microscopic model developed.

Anyway, Figure 57 shows the architecture of the synchronous microscopic infrastructure model realized. As can be seen input data are administered within four different interacting modules:

- *Infrastructure module,*
- *Rolling stock module,*
- *Signalling system module,*
- *Timetable module.*

Moreover a synchronous simulation core, simulates for each time step into which the total simulation period is subdivided, all network events following a chronological order (as they occur in reality) and determining for each time instant the values of status parameters (e.g. aspect of a line-side signal, speed-position coordinates of rail vehicles) for all network components. Then, several kinds of output data can be provided by such model. For example typical outputs can be: train motion diagrams (e.g. time-speed, time-distance trajectories), configuration of train conflicts, power consumption diagrams (since the model has also a module for calculating train energy consumption), or train statistics such as arrival delays, train punctuality, and so on.

Furthermore, it is necessary to specify that at this time, the model realized does not have a GUI (Graphical User Interface), but is available as a pure source code, whose interface with the user is constituted by a simple Win-32 Console window.

In the following paragraphs a more detailed description will be given for each one of the four modules which constitute the model.

4.3. Infrastructure module

Railway network, has been modelled as a link-orientated graph, where therefore nodes contain only information about positions of stations, signals or switches, while for links all rail track characteristics like radii, gradients, speed limits, must be specified. The practical implementation of the graph model has required the specification of both node and link objects. In particular the attributes assigned to nodes are:

- *Node ID* (i.e. an identification number to univocally identify the node),
- *Node spatial coordinates* (X and Y coordinates in Km).

Specifically, this module does not consider the so-called “signal” nodes, i.e. those nodes which specify positions of signals, balises, or other signalling equipments. Here in fact a node must be considered each time that at least one link attribute changes. Link-

orientated graph models in fact makes the assumption that each link must have homogeneous characteristics, therefore each time that one link attribute changes (e.g. a change of speed limit, a change of gradient, etc.) a new homogeneous link must be created and therefore a new node must be considered. Hence, once the real network infrastructure has been subdivided within homogeneous links and their corresponding start and end nodes, attributes of such nodes must be provided here as input data. Moreover, also the specification of station nodes, containing the position at which the platform of each station is located, is required within this module.

Figure 58, shows how a simple double-track railway network can be represented into the link-orientated graph model considered within the infrastructure module. As said before a new graph node must be taken into account each time that at least one rail track attribute changes, or when a station platform is present. In fact, if rail track attributes g , r , v , respectively represent the gradient, the radius and the speed limit of that track section, it is possible to notice that in the graph model (in Figure 58) a new node is inserted when at least one of the values of these attributes changes.

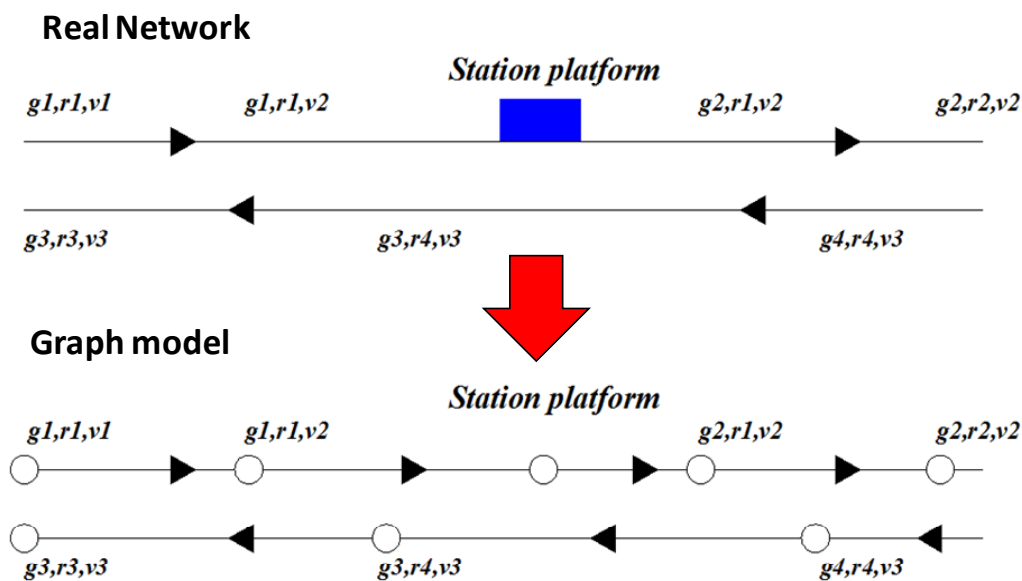


Figure 58. Link-orientated graph model of a simple double-track real network.

Instead, for what concerns graph links, it has been necessary to declare a new class having the following attributes:

- *Link ID* (i.e. an integer number to univocally identify the link)

- *Start node ID* and *End node ID* (i.e. the ID numbers of respectively the begin and end node)
- *Curvature radius* (the rail track section radius in m)
- *Gradient* (positive for uphill and negative for downhill)
- *Speed limit* (maximum civil speed consented on the link in m/s)

The length of each link is automatically calculated by the Pythagorean theorem, since start node and end node positions have already been entered as input. Moreover, direction of links must be specified, in order to provide a rule for train circulation on the network. A Cartesian reference system has been considered for establishing a direction convention for circulation (Figure 59). In particular a value of 1 is given to all links whose running direction is the same as the positive X axis, while a value of -1 is attributed for links whose running direction is the opposite. Then a 0 value is given instead to links which can be crossed in both directions (e.g. single rail track sections).

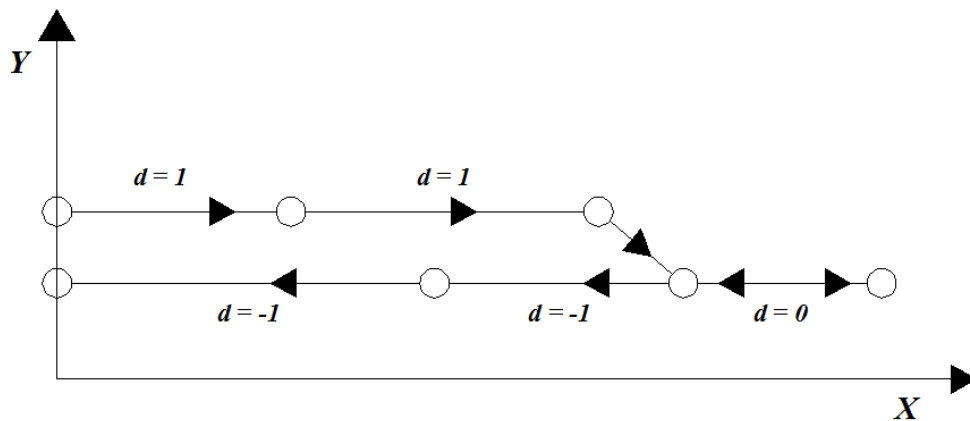


Figure 59. Direction convention for links of the graph model.

Given the convention used within this infrastructure module, all tracks orientated in the same direction as the positive X axis, compose the so-called “Even” track, while all tracks orientated in the opposite direction constitute the “Odd” track. These terms are inherited by the formal Italian terminology for indicating train circulation rules.

By default, the infrastructure module here presented considers a double-track network, therefore the Win-32 console interface will require to enter node and link attributes for both the “even” and the “odd” track. In this case it will be not necessary to specify link

directions, but they will be set automatically by the program. Anyway, it is also possible to skip this default configuration, but it will be required to specify for each link its own direction.

Node and link attributes must be arranged within a database format in a text file. As shown in Figure 60, the fields (i.e. columns) constitute attributes of the elements (nodes or links) while these latter are contained within records of the database. In particular, Figure 60 illustrates the case of a double-track network (considered by default by the program), since link direction is not taken into account within the link database. Instead, if the default mode is skipped, it is necessary to insert a further field in the link database (direction) in which direction of all links (1, 0, or -1) must be reported.

Node Database

Node ID X[km] Y[km]

Link Database

LinkID Start End Radius Gradient Speed

 node ID node ID [m] [m/s]

Node ID	X[km]	Y[km]
1	0.0000	0.000
2	0.1890	0.000
3	0.2180	0.000
4	0.2710	0.000
5	0.3940	0.000
6	0.5140	0.000
7	0.5430	0.000
8	0.6690	0.000
9	1.0140	0.000
10	1.1640	0.000
11	1.3140	0.000
12	1.4600	0.000
13	1.6140	0.000
14	1.8810	0.000
15	2.1510	0.000
16	2.2070	0.000
17	2.3340	0.000
18	2.4210	0.000
19	2.5620	0.000
20	2.6340	0.000
21	2.7180	0.000
22	2.7500	0.000
23	3.0130	0.000
24	3.1490	0.000
25	3.3920	0.000
26	3.5300	0.000
27	3.7280	0.000
28	3.7970	0.000
29	3.8930	0.000
30	3.9622	0.000
31	4.1122	0.000
32	4.2522	0.000
33	4.3220	0.000
34	4.4610	0.000

LinkID	Start node ID	End node ID	Radius [m]	Gradient	Speed [m/s]
100	1	2	10000	0.000	25.00
101	2	3	10000	0.002	25.00
102	3	4	350	0.002	12.50
103	4	5	10000	0.002	25.00
104	5	6	10000	0.005	25.00
105	6	7	10000	0.004	25.00
106	7	8	790	0.004	25.00
107	8	9	790	-0.001	25.00
108	9	10	790	0.000	25.00
109	10	11	790	0.001	25.00
110	11	12	10000	-0.001	25.00
111	12	13	2000	-0.001	25.00
112	13	14	10000	-0.001	25.00
113	14	15	374	-0.001	25.00
114	15	16	10000	-0.001	25.00
115	16	17	600	0.000	20.83
116	17	18	600	-0.002	20.83
117	18	19	600	-0.002	20.83
118	19	20	10000	-0.002	25.00
119	20	21	10000	-0.004	25.00
120	21	22	500	-0.010	19.44
121	22	23	10000	-0.010	25.00
122	23	24	1000	-0.010	25.00
123	24	25	10000	-0.001	25.00
124	25	26	1000	-0.001	25.00
125	26	27	1000	-0.006	25.00
126	27	28	10000	-0.010	25.00
127	28	29	568	-0.014	20.83
128	29	30	10000	-0.014	25.00
129	30	31	546	-0.014	20.83
130	31	32	1111	-0.014	20.83
131	32	33	10000	-0.001	25.00
132	33	34	2200	-0.001	19.44
133	34	35	10000	-0.012	25.00

Figure 60. Input text files of the infrastructure module containing the database of respectively node and link attributes of the “even” track (within the default mode of double-track layout).

Furthermore, to insert such databases as input within the infrastructure module, it is enough to type in apposite spaces indicated by the console interface, the paths of text files containing the data. Figure 61, shows the case of the “Even” track definition. Supposing that the paths of the two text files illustrated in the figure above are

respectively: “C:/TEMP/NodeTrack.txt” and “C:/TEMP/ArcTrack.txt”, it is possible immediately to enter infrastructure input data, by simply typing these path addresses in the console. Then, the same procedure must be repeated for the “odd” track, and once all infrastructure input data are entered, the corresponding graph model is automatically created by the program.

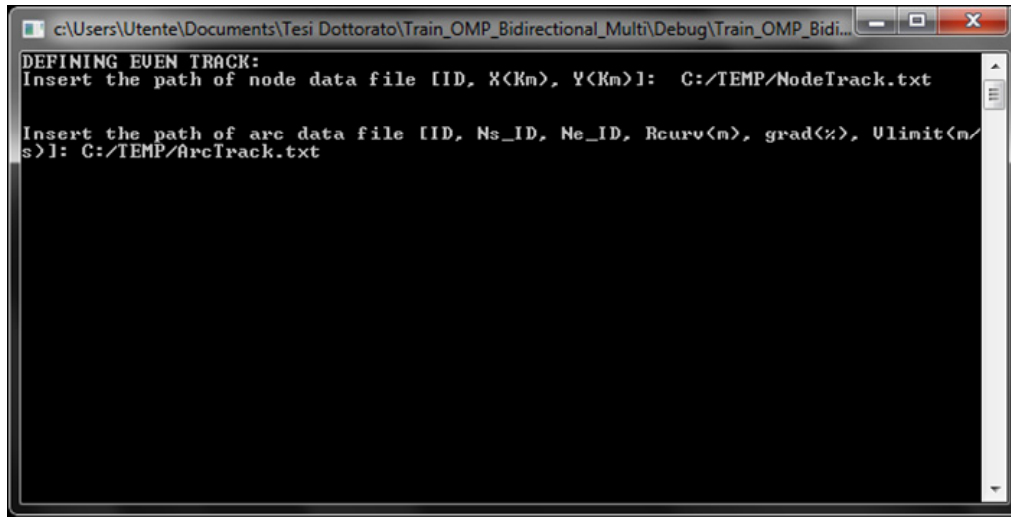


Figure 61. Win-32 Console interface for entering infrastructure input data, within the default configuration of the module (which considers a double-track layout).

4.4. Rolling Stock module

To have an accurate description of rail vehicle dynamics which also considers transient motion phases (e.g. acceleration and deceleration), a detailed depiction of all physical and mechanical vehicle characteristics, has been necessary. To this purpose a “vehicle” object has been declared, and all features such as vehicle length, weight, number of coaches, deceleration rate, etc., must be specified. Moreover, coefficients of the characteristic “tractive effort-speed” curve of the locomotive, as given by equation (13a), must be inserted. Therefore, vehicle attributes that users have to enter within such module are listed below:

- Train *ID* (an integer number to univocally identify trains, which is automatically set by this module according to train departure sequence)
- Mass of the traction unit, m_T [kg]
- Mass of a single wagon, $m_{w,i}$ [kg]
- Number of wagons, n_W

- Maximum speed of the traction unit, v_{max} [m/s]
- Service deceleration rate b_s [m/s²]
- Cross-sectional area of vehicles, A_f [m²]
- Southoff formulae coefficient c_b
- Jerk rate, J [m/s³]
- Total length of the rail vehicle, L [m]

This group of attributes is used to initialize both physical train features as well as parameters of motion resistance equations. Obviously these attributes are fundamental for the integration of the Newton's motion formula, and therefore simulating train movements on the track. As can be seen, the weight of the traction unit m_T is also employed to determine traction unit resistances, that in this module are calculated by using the Italian FS formula, already reported at equation (16), and illustrated again here for convenience:

$$R_{TR}(v) = 4.2 \cdot m_T \cdot g + 0.72 \cdot v^2 .$$

The other attributes, like for example cross-sectional area A_f , the resistance coefficient c_b , as well as the number of wagons n_w and their mass $m_{w,i}$, are important for calculating motion resistances due to air viscosity, as given for example by the Southoff formulae for passenger trains, already reported at equation (17) and here shown again for convenience:

$$R_w(v) = (1.9 + c_b \cdot v) \cdot \frac{g \cdot m_w}{1000} + 4.7 \cdot (n_w + 2.7) \cdot A_f \cdot \left(\frac{v + 15}{10} \right)^2 .$$

The mass of all wagons m_w , contained in such formula, is simply obtained as the product between the number of wagons and the mass of a single wagon: $n_w \cdot m_{w,i}$, if all wagons are of the same type. The length of trains L , is important instead to accurately calculate occupation times of each block section, since clearing and release times depend on both the type of signalling system implemented on the track and train lengths. Therefore in this sense, train length also influence a correct estimation of network capacity.

However, to enter such vehicle attributes within this module, the user must type in the space provided by the Win-32 console interface, the path of a text file within which all this information is contained. For example Figure 62 shows the case in which all these data are collected in a text file whose path is: “C:/TEMP/DataET400.txt”.

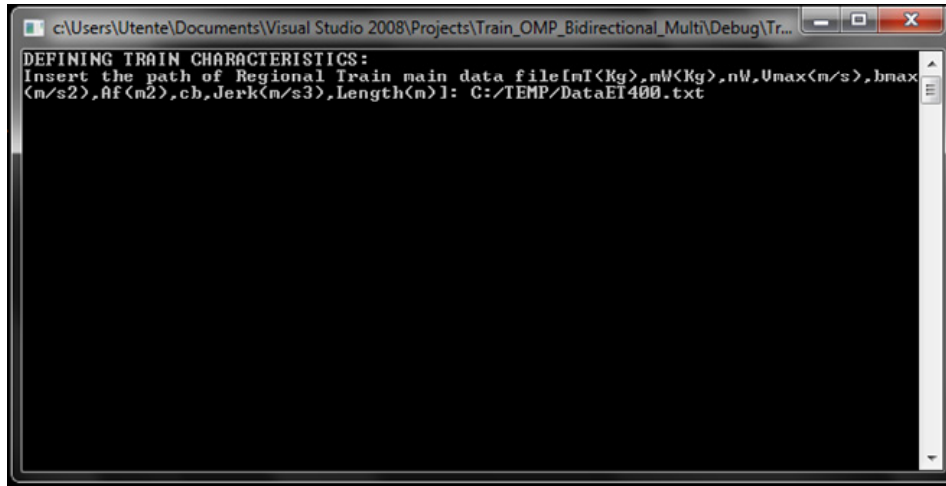


Figure 62. Win-32 console interface to enter train physical parameters within rolling stock module.

Furthermore these data must be arranged within this text file as illustrated by Figure 63.

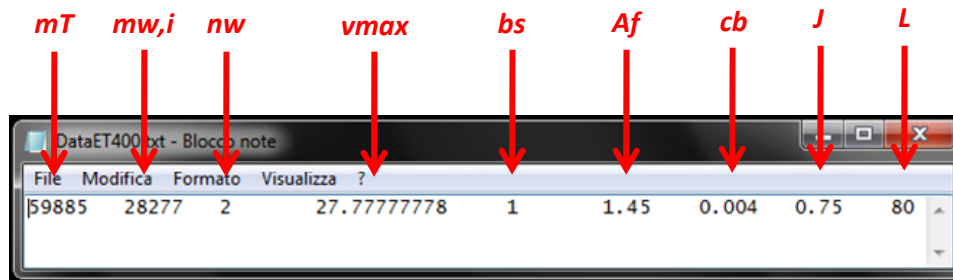


Figure 63. Input text file containing train physical characteristics.

For what concerns instead mechanical features of the traction unit, it is necessary to specify as input data all coefficients of the characteristic “tractive effort-speed” curve relative to the locomotive, given under the mathematical form represented at (13a) and here reported again for convenience:

$$F_{Ti}(v) = c_{0,k} + c_{1,k} \cdot v + c_{2,k} \cdot v^2, \quad v_k < v \leq v_{k+1}$$

As said before, this equation returns the tractive effort between the wheel rim and the rails in function of the instantaneous running speed v . As already said, such curve can be numerically depicted by a set of parabolas as illustrated in Figure 64, and each one of

these parabolic curves is defined within a certain speed domain whose lower and upper bounds are respectively given by v_k and v_{k+1} .

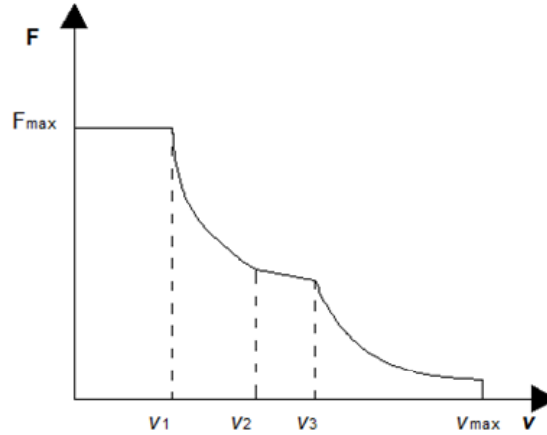
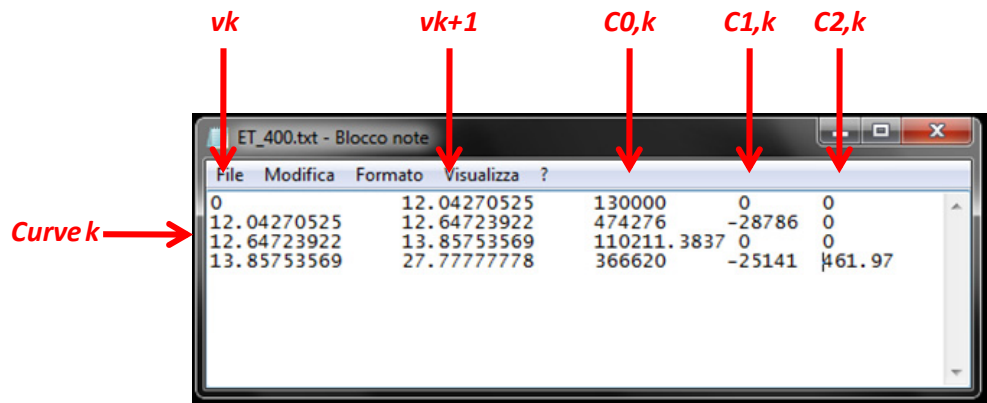


Figure 64. Tractive effort-speed curve of a DC traction unit depicted as a set of parabolas.

Then it is also necessary to specify for each parabola the values of the corresponding coefficients $c_{0,k}$, $c_{1,k}$, and $c_{2,k}$, which determine the shape of the k^{th} parabola. In summary, also the following mechanical attributes of the traction unit must be entered within this module:

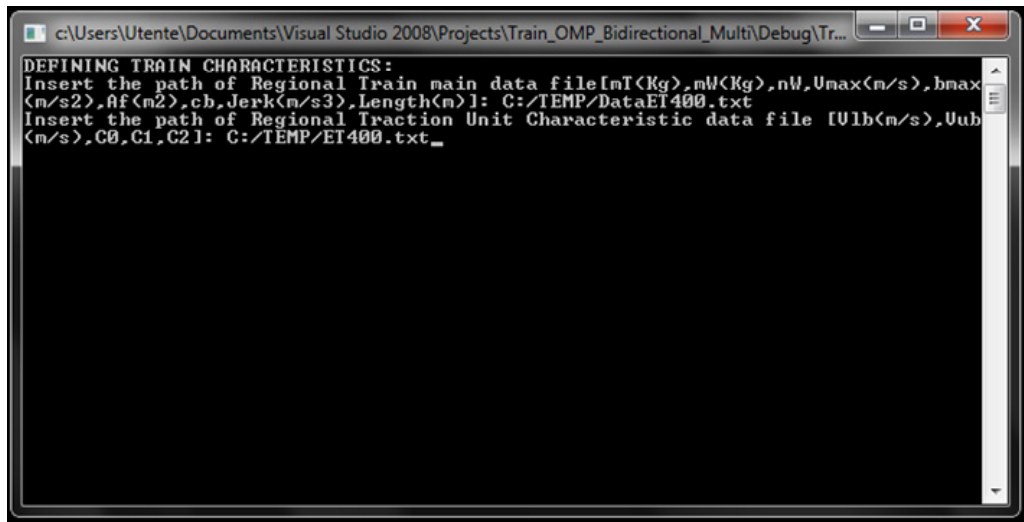
- Lower bound of the k^{th} speed domain, v_k (in m/s)
- Upper bound of the k^{th} speed domain, v_{k+1} (in m/s)
- Parameters of the k^{th} curve, $c_{0,k}$ (N), $c_{1,k}$ (Ns/m), $c_{2,k}$ (Ns²/m²).

In order to enter train mechanical attributes within such module, it is necessary arrange these data in a database format contained within a text file. In particular the k^{th} record of this database will contain both domain speed bounds and coefficients corresponding to the k^{th} parabolic curve through which the mechanical curve has been depicted. While fields of the database (i.e. columns) contain values of domain bounds and curve coefficient. Figure 65 illustrates the database containing the aforementioned data for the case in which the mechanical characteristic curve of the traction unit is represented by four different parabolic curves. Moreover to enter these data, the user must type into the empty row indicated by the Win-32 console interface, the path of the text file into which this database is collected. For instance, Figure 66 shows the described procedure to insert these inputs, supposing that the path of such text file is: "C:/TEMP/ET400.txt".



<i>Curve k</i>	<i>vk</i>	<i>vk+1</i>	<i>C0,k</i>	<i>C1,k</i>	<i>C2,k</i>
0	12.04270525	12.64723922	130000	0	0
12.04270525	12.64723922	13.85753569	474276	-28786	0
12.64723922	13.85753569	27.77777778	110211.3837	0	0
13.85753569	27.77777778		366620	-25141	461.97

Figure 65. Input text files containing mechanical attributes of the traction unit.



```
DEFINING TRAIN CHARACTERISTICS:
Insert the path of Regional Train main data file [mI(Kg),mW(Kg),nW,Unax(n/s),bmax
(m/s2),Af(m2),cb,Jerk(m/s3),Length(m)]: C:/TEMP/DataET400.txt
Insert the path of Regional Traction Unit Characteristic data file [U1b(m/s),Uub
(m/s),C0,C1,C2]: C:/TEMP/ET400.txt_
```

Figure 66. Win-32 console interface to enter mechanical attributes of the traction unit.

Furthermore, the user can also specify the value of adhesion coefficient μ between the wheel rim and the rail, which imposes a different upper bound to tractive effort F , that in this case no longer depends on mechanical features but on adhesion phenomena. In fact the value of tractive effort F used during the integration of Newton's motion formula will always respect the following condition: $F \leq \mu \cdot G$, where G represents the total train weight.

Moreover, different kind of rail vehicles can be considered within the simulation and for each one of them all the aforementioned attributes relative to both physical and mechanical features must be specified by the user. Then for each kind of train, the number of train to simulate which belong to a certain category, can be directly entered by the user through simply modifying a variable (called N_Train) within the program code. In addition, in the same way it is possible to set a regular departure headway

between trains belonging to the same category, through changing the value of the headway variable (called *Headway*) in the program code. In this way a cyclic timetable will be automatically created. Instead when train departures are not regular it is necessary to specify train departure times within the timetable module, as successively described in the dedicated paragraph.

4.5. Signalling system module.

This module models the behaviour of signalling equipments and their interactions with the rail vehicles and other network components. As illustrated in the second chapter of this thesis, signalling systems are the responsible for the safe regulation of train movements on both the open track (e.g. tracks between stations) and within station or shunting areas. This is realized with the aid of signals and other warning items which give movement authorization only if safe conditions for the train are guaranteed. However this module requires the specification of the type of signalling system implemented on the network, the corresponding speed code pattern, the layout of block sections as well as positions and features of switches.

In particular, within this module three different kind of signalling system have been implemented:

- Coded track circuit system (specifically the Italian version B.A.C.C.)
- ETCS level 1 system (the Italian version SCMT)
- ETCS level 2 system.

To accurately model the behaviour of each one of these signalling systems as well as their interactions with other network components a specific class has been declared. Specifically, the user must first specify the kind of signalling system by modifying within the program code the value of a discrete variable (called *Signalling_Level*): if this variable is set to 0, coded track circuit will be used during simulation, instead if it is set to 1 the ETCS level 1 system will be considered, while if it is set to 2 then the ETCS level 2 system will be employed. Moreover, also speed codes relative to different signal aspects of coded track circuit can be established by the user setting the corresponding global variables (*SC1*, *SC2*, *SC3*, etc.) within the program code, expressing their values in m/s. In addition also the delay time of the signalling equipments to communicate the signal aspect to the train is considered. In fact this parameter can be set modifying the

value of a variable (called S_Delay) within the program code, and expressing this value in s.

Successively, it is necessary to define positions of signals and/or balises, and therefore the layout of block sections. In particular, once the type of signalling system is defined, the user has to specify block section lengths and after that, the corresponding signalling layout will be automatically created. The graph model representing the infrastructure network (and defined within the infrastructure module) will be in fact enriched with additional signal nodes positioned at block section joints as defined by the user. Therefore the signalling structure is laid upon the physical infrastructure network, and the resulting network graph is composed of both physical (e.g. stations) and signal nodes (e.g. line-side signals), as shown in Figure 67.

Network graph model

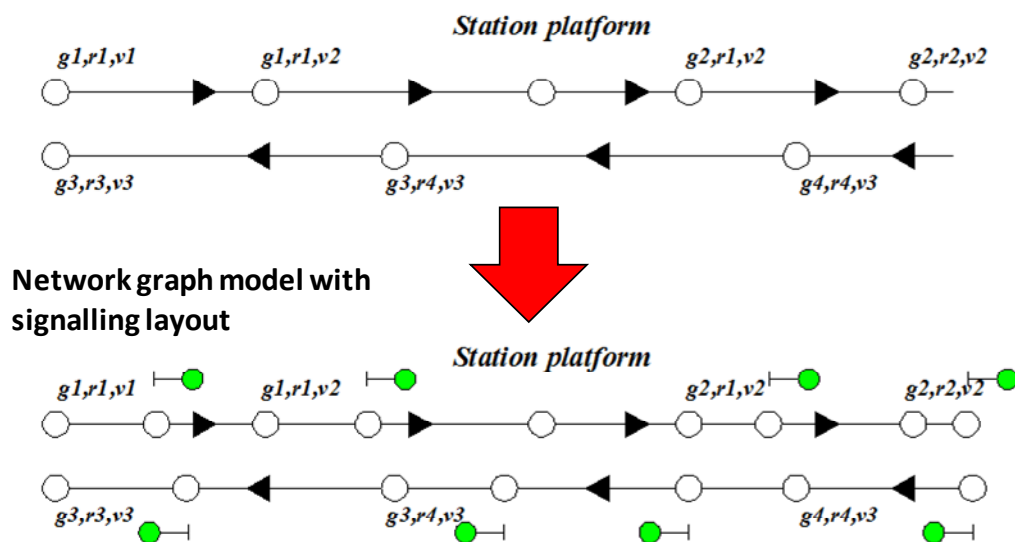


Figure 67. Graph model of the network before and after the definition of the signalling layout.

Obviously the length of each block section will define the distance between two consecutive main signals. Then two different alternatives are available by the user, in fact it is possible to define: 1) an equi-block section layout, in which each block section have the same length, or 2) a non-equal block section layout, where block sections can have instead different lengths. The implementation of the first option (equi-block layout) is very simple, and it is enough for the user to specify in Km the length of the equi-block layout, through setting the corresponding value of a variable (called $EquiBlock_length$) in the program code. In this way all block sections belonging to

station areas will be automatically created according to an equi-block layout, while on open tracks, block sections can assume also different lengths with respect to the established equi-block length. This is due to the fact that not always the ratio between the length of an inter-station track L_i and the equi-block length EBL , is an integer value. In these cases in fact the remaining length must be partitioned in the same amount for all the sections which compose that open track. For example, if a certain inter-station track has a length of $L_i = 3.2$ Km and the equi-block length imposed is equal to $EBL = 0.564$ Km, it is immediate to understand that their ratio is not an integer value $L_i/EBL = 5.67$. Therefore the 5 equi-block sections long 0.564 Km, must be stretched of a total of $(3.2 - 5 \cdot 0.564) = 0.38$ Km. Hence each one of these 5 block section must be stretched of 0.076 Km to reach a length of 0.64 Km.

The implementation of the second block section layout, requires instead the specification of lengths for each block section. This can be realized arranging block section data within a database, which can be contained in a text file format. In particular the attributes that users must define are the ones listed below:

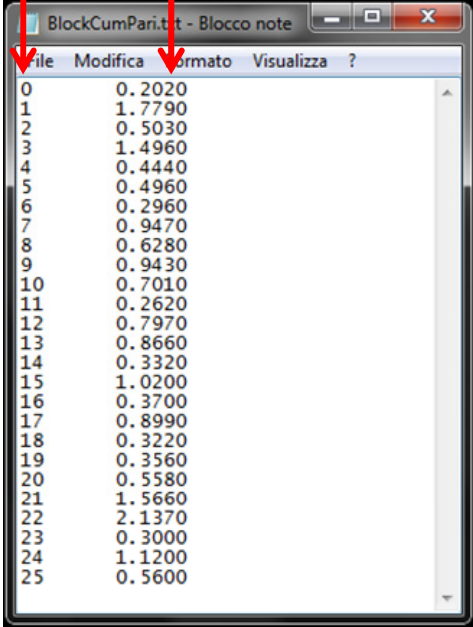
- *Block section index* (i.e. an integer value starting from 0 to n according to the sequence that block sections have following the positive X axis direction)
- *Block Section length* (the length of each block section in Km).

In the case of networks with double-track layout it is necessary to define these data for both the directions. Anyway, Figure 68 illustrates the text file containing the attributes of block section layouts. Then, to enter such data within the signalling system module, it is necessary to type the path of this text file within the blank row as indicated by the Win-32 console interface, which is shown in Figure 69. Therefore, supposing that the path of this text file is: "C:/TEMP/BlockCumPari.txt", it will be enough to type this address to insert the corresponding block section layout.

Another fundamental part of the signalling system module, concerns instead the definition of interlocking system and in particular points and switches, used to set train paths on the network. As illustrated in the second chapter, switches consent the safe management of train paths both within complex areas (e.g. stations with many platforms, marshalling yards), and railway junctions, allowing trains to change tracks

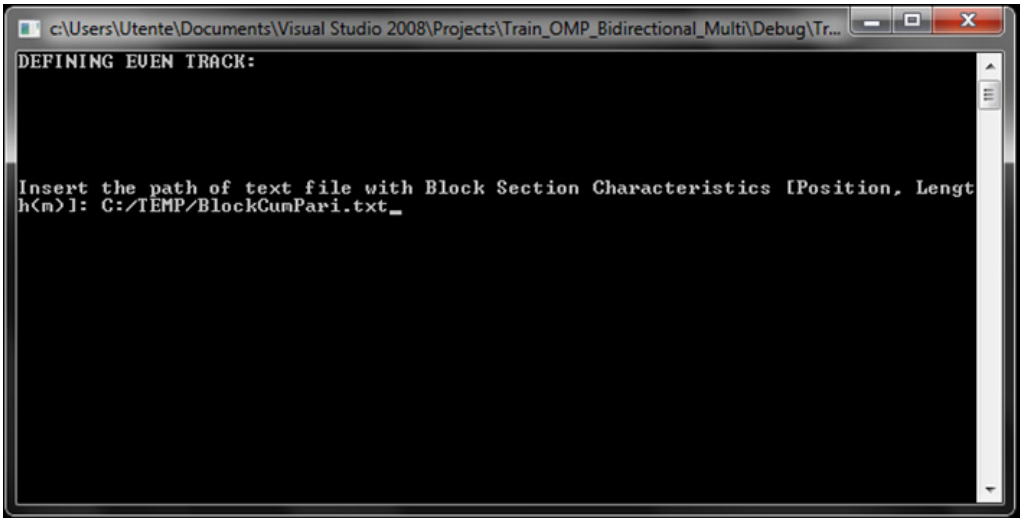
according to the scheduled service or to re-routed paths indicated by dispatching decisions to solve real-time conflicts.

Block Index *Block length*



0	0.2020
1	1.7790
2	0.5030
3	1.4960
4	0.4440
5	0.4960
6	0.2960
7	0.9470
8	0.6280
9	0.9430
10	0.7010
11	0.2620
12	0.7970
13	0.8660
14	0.3320
15	1.0200
16	0.3700
17	0.8990
18	0.3220
19	0.3560
20	0.5580
21	1.5660
22	2.1370
23	0.3000
24	1.1200
25	0.5600

Figure 68. Input database for the definition of block section layout.



```
DEFINING EVEN TRACK:

Insert the path of text file with Block Section Characteristics [Position, Length]
h<n>l: C:/TEMP/BlockCumPari.txt_
```

Figure 69. Win-32 Console interface for entering the database relative to block section layout.

In order to guarantee safe train movements, interlocking systems allow a train to proceed on a certain switch in a diverging position, only if safe conditions are verified. To realize these conditions a restricted aspect is given to all potential conflicting trains. Therefore, the correct description of interlocking systems and in particular network

switches requires a detailed modelling of switches behaviour and above all their interactions with network signals and rail vehicles. In general, a switch connects two different tracks and therefore its position (diverged or not) conditions the aspects of signals relative to the involved block sections. The example represented in the case A of Figure 70, clearly shows that when switch SW1 is locked in the diverging position to allow the movement of train T1, main signals of block section BS3 show a red aspect to avoid dangerous opposite movements of train T2. Therefore in this condition when the switch SW1 is locked in the diverging position, it consents to train T1 to enter the block section BS2, blocking at the same time the access at block section BS3, to opposing train T2. On the contrary, when the switch SW1 is not locked in the diverging position (case B), it allows to train T2 to enter block section BS3, while train T1 must respect the restricted aspect given to main signals of block section BS2. As can be easily understood, to accurately model the behaviour of interlocking systems and therefore the dependence between signals and moving elements (i.e. switches), it is necessary to define for each position of the switch, which block section shows a proceed aspect and which instead is blocked.

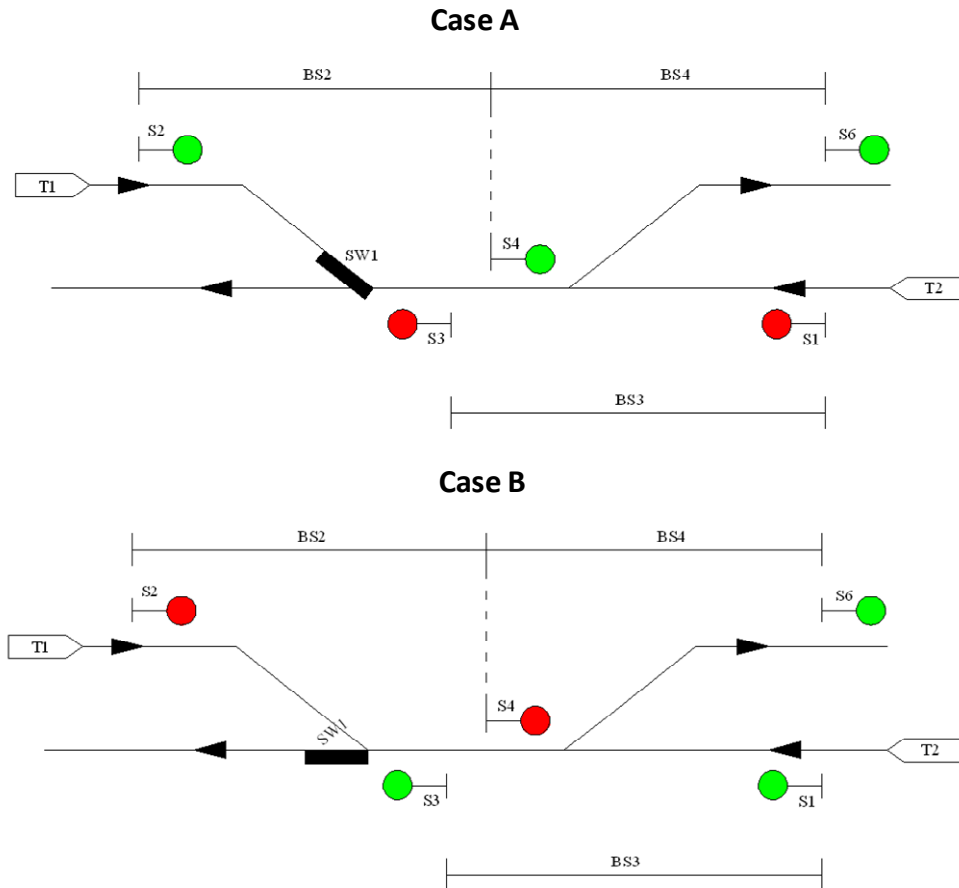


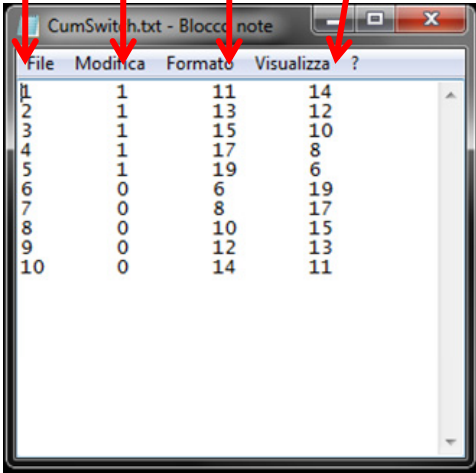
Figure 70. Example of an interlocking system, case A: a switch locked in a diverging position, case B: a switch locked in the non-diverging position.

In such module, in fact the position of movable elements is indicated through a Boolean variable which is equal to 1 when the switch is diverged and to 0 if it is in the standard position. Then it is necessary to specify the indexes of both the block section set to a proceed aspect and the block section set instead to a red aspect. Therefore, considering the case A shown in Figure 70, the user must explicitly define that when the switch SW1 assumes position 1 (diverging), it is possible to access at block section BS2, while section BS3 is blocked to opposing trains. Therefore, this module requires the definition of further attributes addressed to determine such interlocking components. In particular these attributes are:

- *Switch ID* (an integer number which univocally identifies the element)
- *Position* of the switch (1 to indicate diverging position and 0 for standard position)
- *Index of the block section* which shows a proceed aspect when the switch is locked in the position specified by “Position” attribute.

- *Index of the block section* which shows a restricted aspect when the switch is locked in the position specified by “Position” attribute.

Figure 71 illustrates, the database containing the aforementioned attributes to define all switches on the track. Considering as example the first row of this database it is possible to see, that when switch 1 (i.e. with ID = 1), is locked on the diverging position (i.e. Position =1), the block section 11 shows a proceed aspect while block section 14 is blocked.

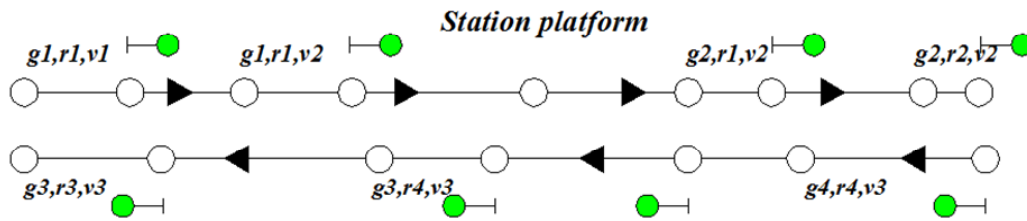


Switch ID	Switch position	Block index (proceed)	Block index (blocked)
1	1	11	14
2	1	13	12
3	1	15	10
4	1	17	8
5	1	19	6
6	0	6	19
7	0	8	17
8	0	10	15
9	0	12	13
10	0	14	11

Figure 71. Input database for the definition of the layout of interlocking elements.

When such data are entered within the signalling module, switches are automatically created within the graph model of the infrastructure network, as illustrated in Figure 72.

**Network model with
signalling layout**



**Network model with signalling and
interlocking layout**

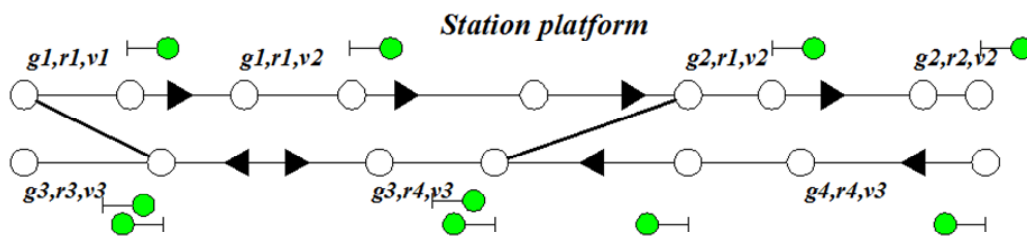


Figure 72. Graph model of the network before and after the specification of the interlocking layout.

However, these input data can be entered in this module typing in the apposite blanks indicated by the Win-32 console interface, the path of the text file in which they are contained. Figure 73 shows for example the case in which the path of this text file is: “C:/TEMP/CumSwitch.txt”.

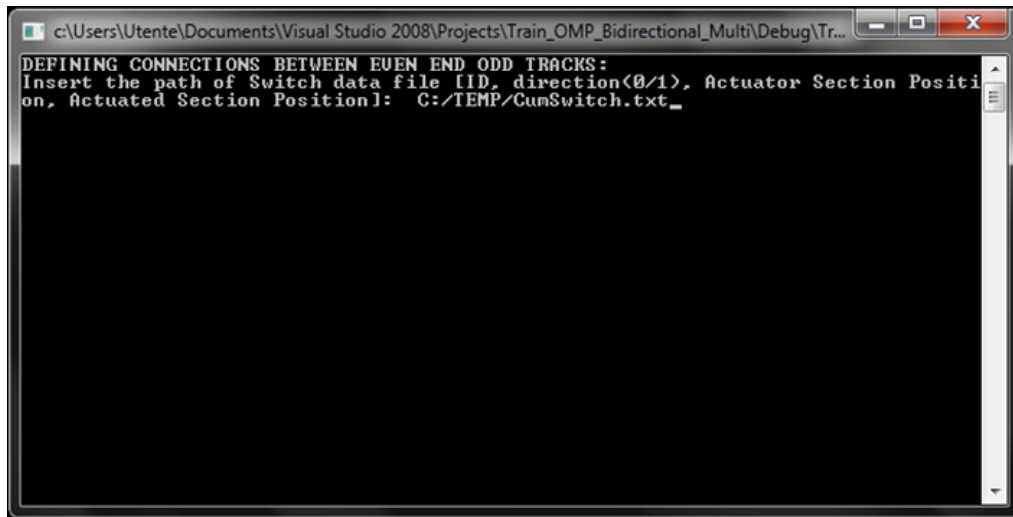


Figure 73. Win-32 console interface for entering data relative to the interlocking layout.

4.6. Timetable module

Timetable dictates and regulates train operation within the network. It establishes in fact train paths, platforms at which train has to stop within stations, and above all departure times, arrival times as well as dwell times of trains at stations. This module is therefore fundamental, since it defines the chronological sequence according to network events must be processed. Moreover, train performances in terms of punctuality or average delay are just commensurate with respect to scheduled arrivals or departures at/from key network points, as defined by the timetable itself. Hence, such module needs to take as input data train departure/arrival times at each station, or if a regular service is provided (e.g. for a mass rapid transit line), it is enough to specify train headways and dwell times at stations.

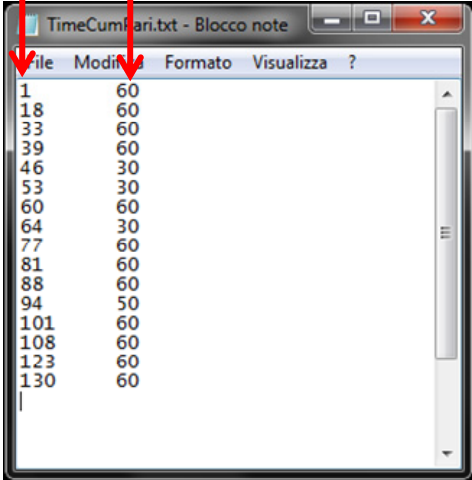
In such module, timetable is represented as a simple database containing all information about scheduled train operations. Therefore, each train object (defined within the rolling stock module) read its corresponding schedule of operation from this database.

In particular, for systems with homogeneous and regular traffic such as mass rapid transit lines, the user must define train headway, setting the value (in seconds) of a global variable (called *Headway*) within the program code. Then the definition of train dwell times at each station is necessary to automatically create a cyclic timetable. Specifically, to define a similar kind of timetable, dwell times must be defined (in seconds) for each station, arranging these information in a database whose fields are just constituted by:

- *Station Node ID* (i.e. the identification number of nodes representing stations, this ID must be taken by the Node database introduced in the infrastructure module),
- *Dwell time* (i.e. train dwell time at that station, expressed in seconds).

Figure 74 illustrates the database containing the aforementioned attributes necessary to define a cyclic timetable. Each record specifies these attributes for a certain line station. Then to practically enter these data in this module, the path of the text file within which they are contained must be typed in the blanks indicated by the Win-32 console interface. For instance, Figure 75 shows the case in which the path name of this text file is: "C:/TEMP/TimeCumPari.txt".

Station Dwell
Node ID time



1	60
18	60
33	60
39	60
46	30
53	30
60	60
64	30
77	60
81	60
88	60
94	50
101	60
108	60
123	60
130	60

Figure 74. Input database for the definition of a cyclic timetable.

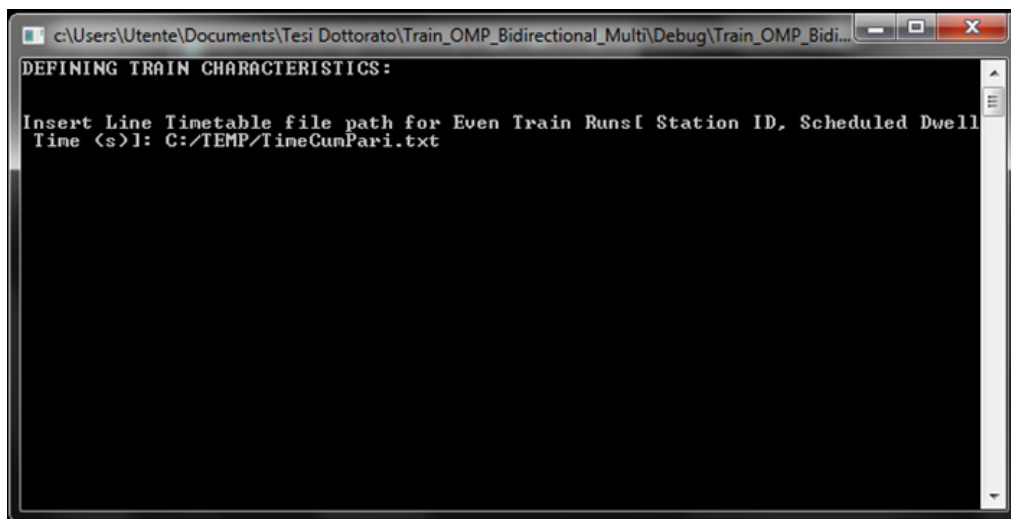


Figure 75. Win-32 console interface for entering data relative to timetable.

For double-track lines, with trains serving both directions, it is necessary to specify the timetable for all trains involved, and therefore for both service directions. In this case two databases must be considered, one for each direction. Hence, the procedure described above for entering timetable data, must be repeated for both service directions.

Furthermore, this module consents to take into account stochastic disturbances to train operations, giving the possibility of considering train dwell times at station as random variables with a certain probability density function. In particular three different density probability functions are available:

- *Normal distribution*
- *Negative exponential distribution*
- *Log-normal distribution*

Usually the negative exponential is used to model departure delays of trains, as shown in literature by Schwanhäußer (1974) and Yuan (2006), in fact to this purpose also train departure delays can be assumed in this module, as a negative exponential random variable. For each type of these distribution functions, users has to specify the values of both the average and standard deviation parameters. This can be simply done, directly setting the values of input variables of corresponding global functions within the program code. In particular, the function that must be used to define a Normal distribution is *Draw_Normal_Dwell_Time* (double *Average_Dwell*, double *Dwell_StD*). Hence, setting the values in seconds of the input variables *Average_Dwell* and *Dwell_StD*, dwell times of all trains at stations will be distributed as a normal variable with that average and that standard deviation. To model train departure times as negative exponential variables, the function *Draw_Departure_Delay* (double *Average_Delay*), must be employed. However, in this case it is enough to set only the input variable *Average_Delay*, since for a negative exponential distribution, standard deviation is equal to the average. To model dwell times instead according to a log-normal distribution, it is necessary to set values of input parameters of the function *Draw_Log_Normal_Dwell_Time* (double *Average_Dwell*, double *Dwell_StD*).

Moreover, it is also possible to set for each station a different value of the average dwell time, which can be for instance constituted by the dwell times specified in the timetable database illustrated in Figure 74. Stochastic modelling of dwell times, is fundamental when analyzing stability or robustness of timetables or network infrastructure, as well as investigating on network performances and the effectiveness of certain operation strategies under disturbed operations.

4.7. Simulation core

In this section, the simulation process used by the developed microscopic model is described. Once all input data have been inserted within each one of the four modules depicted in the previous paragraphs, both infrastructure network (including signalling and interlocking systems) and trains (including vehicles and operational timetable) are

defined. Since it is a synchronous microscopic model, all events are simulated in the same sequence that they have in reality. However, before launching the simulation it is necessary to set within the program code the value of two global variables: the time step (called just *timestep*) expressed in seconds, as well as the simulation period (called *times*) expressed in number of time steps from which it is composed of. For example if the time step is assumed equal to 1 second, to set one hour simulation period, the *times* variable must be set to 3600. Instead if the time step is assumed equal to 0.1 seconds, the variable *times* must be set equal to 36000, since 1 hour is constituted of 36000 tenths of second.

For the sake of clarity, a time-discrete simulation is here considered. That means that the simulation clock goes ahead with discrete time where each time instant t is obtained as the sum of the previous time instant $t-1$ and the defined time step Δt : $t = t-1 + \Delta t$.

Moreover it is also possible to perform different replications of the scenario, each one with a different random seed (for drawing stochastic disturbances to train operations). The user can in fact define the number of replication setting the value of the variable *N_Replication* directly in the program code. However at the end of each replication the corresponding output are given, while at the end of the whole simulation experiment, results are returned as the average of the outputs from the considered number of replications.

Anyway, when simulation clock is turned on, for each time instant t of the considered period, railway operations are simulated following these steps:

1. For all trains, whose entry time in the network is lower than instant t (therefore for all trains that are on the network at instant t), their position $s[t]$ and speed $v[t]$ at instant t , is calculated by integrating the Newton's motion formula, according to a difference equation approach. In particular, for each time step, the maximum force F_{Ti} between the traction unit's wheels and the tracks is calculated, considering the characteristic curve set as input within the rolling stock module as well as the speed value of the train at instant $t-1$, $v[t-1]$:

$$F_{Ti}(v[t-1]) = c_{0,k} + c_{1,k} \cdot v[t-1] + c_{2,k} \cdot v[t-1]^2,$$

Then motion resistances (for the traction unit, wagons, line gradient and curvature) are calculated for all trains at instant t using the relation already presented in chapter 2: $(R_{TR}(v[t-1]) + R_W(v[t-1]) + R_g + R_c)$.

After that, the speed of each train at instant t , $v[t]$, is calculated integrating the Newton's motion formula:

$$v[t] = v[t-1] + \frac{F_{Ti}(v[t-1]) - (R_{TR}(v[t-1]) + R_W(v[t-1]) + R_g + R_c)}{f_\rho \cdot m} \cdot \Delta t;$$

Successively after that the speed value $v[t]$ has been obtained, train position at instant t is calculated as:

$$s[t] = s[t-1] + \frac{f_\rho \cdot m \cdot (v[t] - v[t-1]) \cdot v[t-1]}{F_{Ti}(v[t-1]) - (R_{TR}(v[t-1]) + R_W(v[t-1]) + R_g + R_c)};$$

In addition, particular attention must be paid for the deceleration phase and the calculation of the corresponding braking curve. In fact in this case, a negative force is applied to train wheels and therefore in the equations reported above the effort F_{Ti} has the same direction as motion resistances (therefore opposite to train running direction) and in particular is calculated as: $F_{Ti} = f_\rho \cdot m \cdot b_s$, where b_s represents the deceleration rate of the train as specified in input within the rolling stock module. Moreover, braking curve is here calculated also taking into account the variation of the deceleration rate during time, i.e. the so-called Jerk, which in fact constitutes one of the inputs of the rolling stock module. In fact, the deceleration rate (and therefore the braking effort) considered when calculating braking distance has the trend shown in Figure 76.

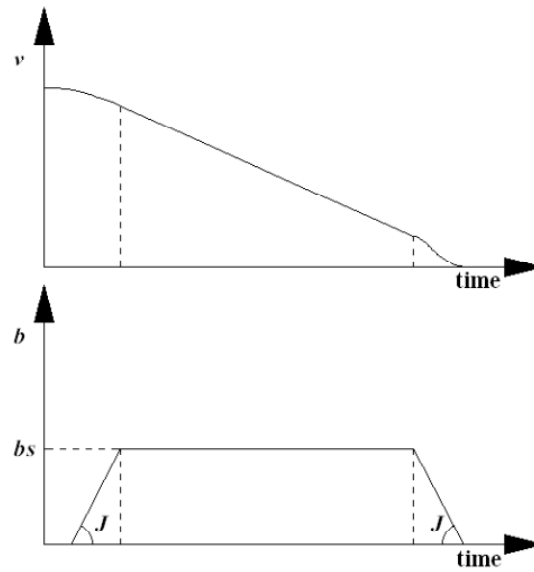


Figure 76. Train braking performance with allowance for jerk control.

Moreover, the braking curve is integrated in a reverse way starting from the objective speed-distance point in order to know at each time instant where the train has to precisely start braking to arrive at the objective point respecting certain speed constraints that can be imposed by both fixed track characteristics (e.g. presence of stations, reduction of civil speed limits) and/or variable information (e.g. signal aspects, switch positions). In fact at each time instant the next “objective speed-distance” point (e.g. the stop at a station, the stop at a red signal, or simply a lower speed limit to observe) is determined for each train, considering both fixed track characteristics as well as the state of signalling system at that instant. Starting from this point and taking into account the current speed and position of the train, the corresponding braking curve is calculated and the point P where the train must start braking is also determined. Therefore, the train begins its deceleration phase when its current position $s[t]$ is: $s[t] \geq P$ (Figure 77).

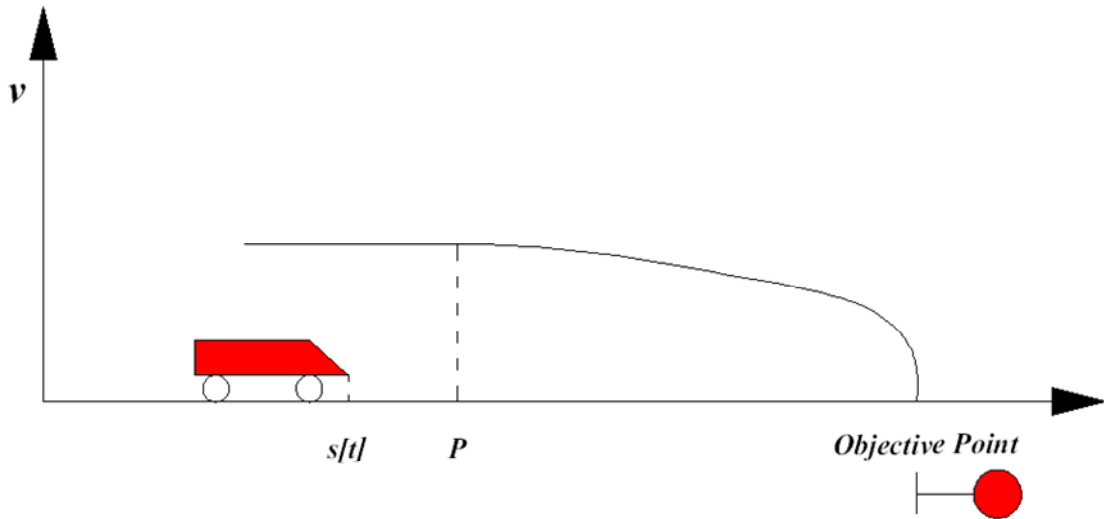


Figure 77. Reverse integration of the braking curve starting from the objective point and determination of the point P where the train starts braking.

Furthermore also mechanical power consumption of trains is here calculated for time instant t . This is immediately carried out through applying the following relationship: $P[t] = F_{Ti}(v[t]) \cdot v[t]$, and therefore estimating the product between the tractive or braking effort (it depends if the train is accelerating or decelerating) and the speed that train has at instant t . Then integrating according to a difference equation approach, the power consumption, mechanical train energy consumed during running is also determined. In fact, the variation of energy consumption between time instants t and $t-1$ can be simply calculated as:

$$\Delta E[t] = (P[t] - P[t-1]) \cdot \Delta t / 2 ;$$

Obviously, the total energy consumed by a train over all the simulation period considered is given by the sum of $\Delta E[t]$ over all time steps composing the simulation period itself.

2. After that speeds and positions at instant t have been determined for all trains on the network, both the signalling and the interlocking systems are updated. In fact the aspects of signals and the configuration of interlocking elements change according to the positions of trains on the line (and in particular to which block section is occupied), and obviously to the type of signalling system considered. For example (Figure 78) if a traditional multi-aspect system (based on track circuits) is implemented and train T1 occupies block section BS2, signalling system will be updated emulating the real aspect sequence of that kind of

system. Therefore section BS3 will show a red aspect, BS4 a yellow aspect, while BS5 a proceed aspect.

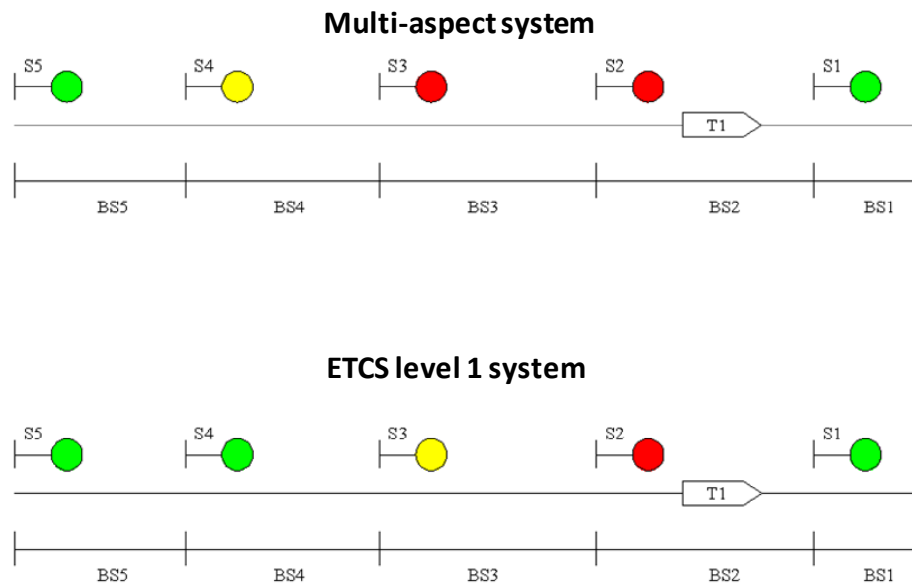


Figure 78. Signalling system updating according to the type of system and train positions

Instead, if an ETCS level 1 system is installed, when train T1 occupies block section BS2 only the preceding section BS3 shows a red aspect while other sections have a proceed aspect.

Therefore aspects of signals are updated according to the current position of trains on block sections, in order to regulate movements on the track emulating real signalling behaviour and obviously respecting safety conditions.

3. Once that both vehicle parameters (speed, position, power and energy consumed) and the configuration of signals and switches have been calculated for time instant t , the simulation clock goes ahead at instant $t+1$ and the described cycle is restarted again from step 1. Obviously, this cycle stops when the whole simulation period has been simulated. At this point output data are returned for each train (e.g. speed-time trajectories, time-distance trajectories, arrival delay) and each station (average delay at station, punctuality at station) in a text file format (simulation outputs will be better described in the following section).

However, observing the scheme illustrated in Figure 79, it is possible to summarize the main steps through which railway operations are simulated by the developed microscopic model.

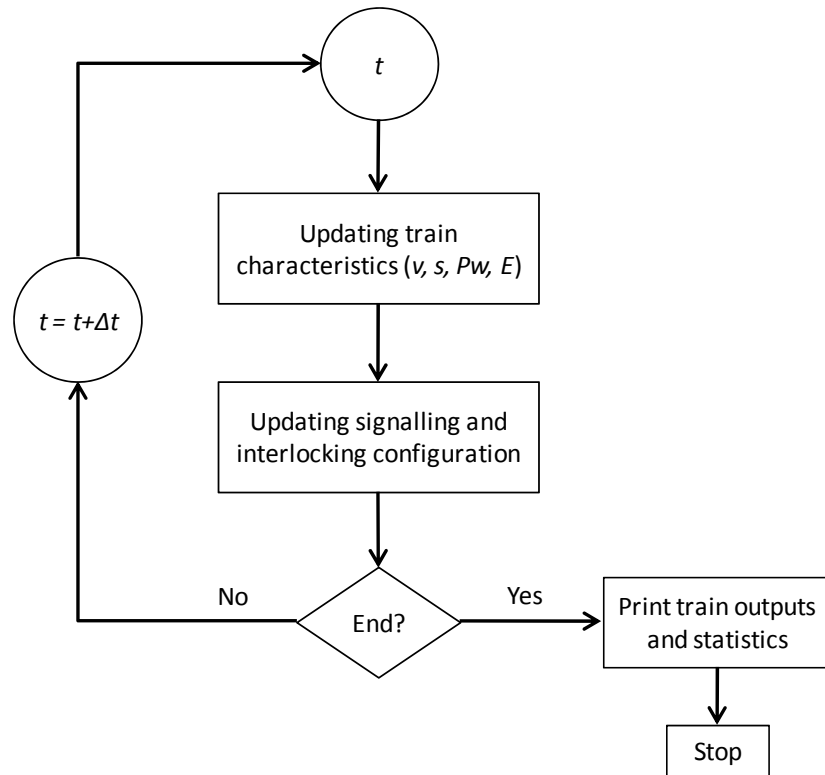


Figure 79. Simulation process of the developed railway microscopic model for a single replication.

4.8. Simulation Outputs

As already said within previous sections, the outputs returned by the developed microscopic model concern train diagrams (e.g. speed-distance trajectories, time-distance diagrams) and statistics (e.g. train arrival delays, punctuality at stations). Moreover diagrams relative to both mechanical power and energy consumed during train runs are given as output, and therefore also information about electrical energy supply needed to operate the system under a certain timetable, is available. In particular when simulation process stops, for each simulated train a text file is printed, containing all information about the dynamic evolution of that train's state during the whole simulation period. Specifically, train outputs are arranged within a database whose fields are respectively constituted by:

- Simulation time instant, $t[s]$

- Train speed, $v[\text{m/s}]$
- Nose position of the train, $s[\text{m}]$
- Tail position of the train, $st[\text{m}]$ (which is equal to the difference between nose position and train length)
- Mechanical power consumption, $P_w [\text{KW}]$
- Mechanical Energy consumption, $E[\text{MJ}]$

Records of this database instead, contain values assumed by each one of the aforementioned fields for each time instant of the whole simulation period. An example of the output database produced by the simulation model is illustrated in Figure 80. In particular, text files containing train output data are printed for each replication of the whole simulation experiment. Moreover, also an output text file is printed containing train arrival times at each station or key network point (e.g. important junctions). Such output is also returned as a database where each column represent a train, while on the i^{th} row it is possible to read the arrival time (in s) of that train at station i (see Figure 81). This is useful for elaborating train statistics on delays or estimating train punctuality indexes. Such file is produced for each replication of the simulation experiment. In addition, it is also possible to have statistics on train arrival delays (e.g. average delay) at a certain station, through simply setting the value of the input parameter *ST_Index* (which is the ID of the station node) within one of the following code functions:

t[s] *v* *s[m]* *st[m]* *Pw[KW]* *E[MJ]*
[m/s]

Time[s]	Speed[m/s]	Position[m]	Tail_Position[m]	Power_Cons[kW]	En[MJ]
0	0	0	-80	0	0
1	0.992643	0	-80	0	0
2	1.98422	0.992643	-79.0074	129.044	0.064522
3	2.97452	2.97686	-77.0231	257.949	0.2380185
4	3.96332	5.95138	-74.0486	386.687	0.5803365
5	4.95043	9.9147	-70.0853	515.232	1.031296
6	5.93563	14.8651	-65.1349	643.556	1.61069
7	6.91871	20.8008	-59.1992	771.631	2.3182835
8	7.89947	27.7195	-52.2805	899.432	3.153815
9	8.87771	35.6189	-44.3811	1026.93	4.116996
10	9.85322	44.4966	-35.5034	1154.1	5.207511
11	10.8258	54.3499	-25.6501	1280.92	6.425021
12	11.7953	65.1757	-14.8243	1407.35	7.769156
13	12.7614	76.9709	-3.02907	1533.39	9.239526
14	13.566	89.7324	9.73235	1406.45	10.709446
15	14.3676	103.298	23.2984	1495.13	12.160236
16	15.0905	117.666	37.666	1447.79	13.631696
17	15.7439	132.756	52.7565	1394.83	15.053006
18	16.3375	148.5	68.5003	1343.14	16.421991
19	16.8796	164.838	84.8379	1293.68	17.740401
20	17.3769	181.718	101.718	1246.98	19.010731
21	17.8348	199.094	119.094	1203.22	20.235831
22	18.258	216.929	136.929	1162.45	21.418666
23	18.6506	235.187	155.187	1124.61	22.562196
24	19.0159	253.838	173.838	1089.55	23.669276
25	19.3567	272.854	192.854	1057.13	24.742616

Figure 80. Text file containing the database of simulation outputs relative to a certain train for a single replication.

Train 1 *Train 2* *Train 3* . . . *Train j*

Train_1	Train_2	Train_3	Train_4	Train_5	Train_6
0	0	0	0	0	0
198	798	1398	108	708	1308
369	969	1569	308	908	1508
493	1093	1693	473	1073	1673
624	1224	1824	605	1205	1805
711	1311	1911	788	1388	1988
808	1408	2008	898	1498	2098
924	1524	2124	997	1597	2197
1080	1680	2280	1190	1790	2390
1206	1806	2406	1305	1905	2505
1329	1929	2529	1434	2034	2634
1476	2076	2676	1529	2129	2729
1604	2204	2804	1631	2231	2831
1771	2371	2971	1756	2356	2956
1969	2569	3169	1930	2530	3130
2131	2731	3331	2120	2720	3320
2206	2806	3406	2181	2781	3381

Figure 81. Text file containing the database of train arrival times at each station for a single replication.

- Calculate_Total_Delay_At_Station (int *ST_Index*), which returns the sum of arrival delays over all simulated trains at the considered station (in seconds).
- Calculate_Average_Delay_At_Station (int *ST_Index*), which gives the average of train arrival delays at the considered station (in seconds).

Moreover such output values can be referred both to each single replication and to the whole simulation experiment as the average result on replications. Anyway, all output text files are saved in a folder named TEMP, which has the following pathname: “C:/TEMP/”.

These text files can be easily imported in software for data analysis (e.g. Excel, Matlab) to immediately elaborate train diagrams such as time-distance trajectories, speed-distance trajectories, power-distance diagrams, or to visualize effects due to train conflicts.

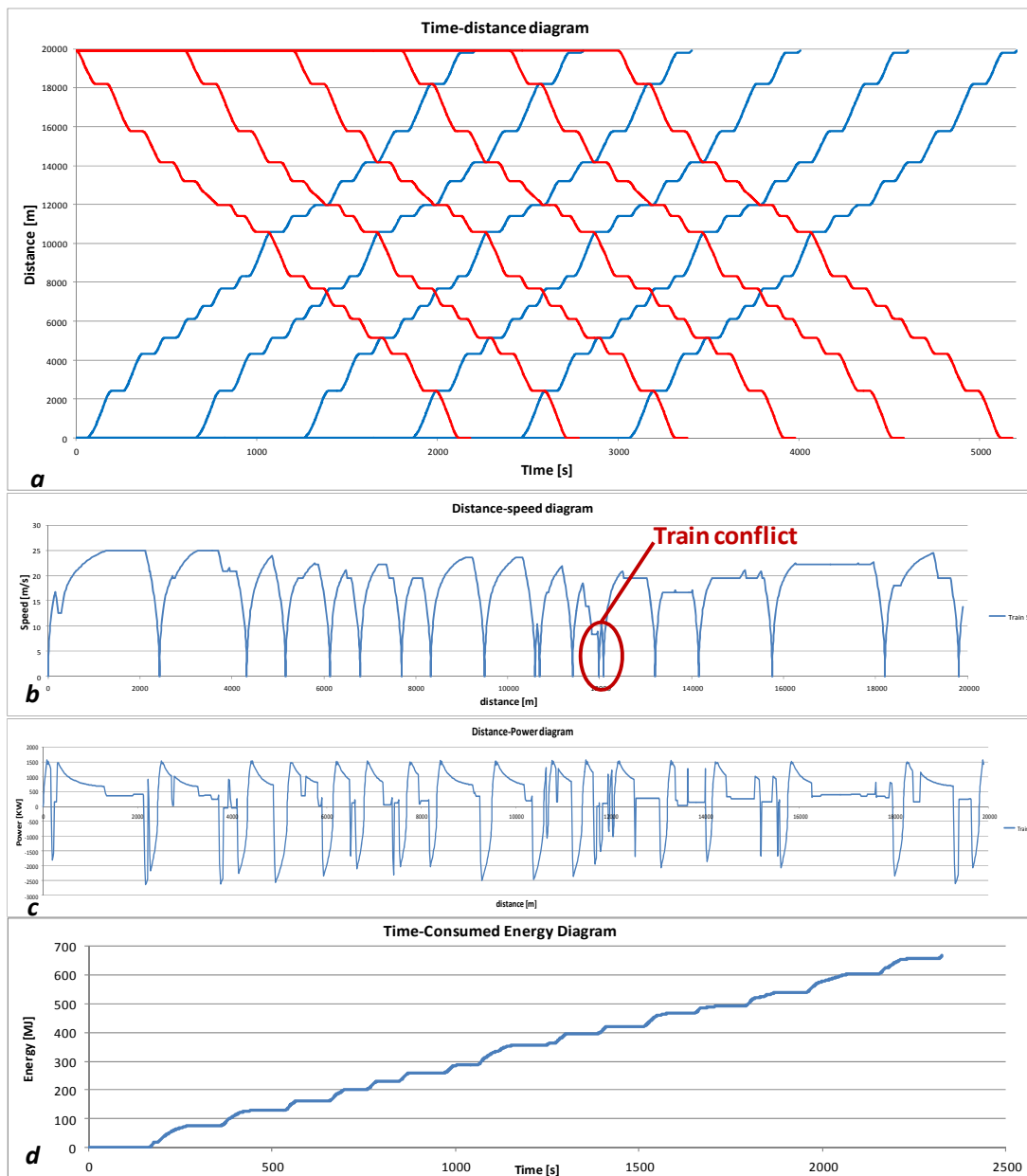


Figure 82. Graphical representation of simulation outputs: a) time-distance diagram, b) distance-speed trajectory for a single train, while in red a forced stop due to a conflict is highlighted, c) distance-power graph relative to a single train, d) time-consumed energy diagram for a single train.

In fact Figure 82 shows some graphical representation of the described output data returned by the simulation model developed. In particular in Figure 82a, time-distance trajectories of trains running on a double-track network are represented, in Figure 82b instead the distance-speed trajectory relative to a single train is reported. Furthermore, here it is also highlighted with a red circle a deceleration phase for stopping at a red signal, since a conflict with the preceding train has arisen. Then, Figure 82c illustrates the distance-power diagram relative to a single train, while in Figure 82d the corresponding curve of energy consumed by train during time has been reported.

However, since the model has an open structure, it is possible to define customized functions in order to obtain any kind of train output, measure of performance or statistic.

4.9. Parallelization and its effects on computing efficiency

The first version of the microscopic model was obviously based on a classical serial architecture. This means that all simulation processes as well as program functions addressed to simulate train operations for a certain time instant, were executed in a serial way, i.e. one could be launched only if the preceding one had been completed. As can be easily understood, this serial execution induced computing times of the simulation process to become higher and higher with the increasing of network dimensions or the congestion level (number of trains on the network), bringing therefore to unreasonable simulation times in these cases. Moreover this kind of architecture did not consent to exploit the power of modern multi-core computers, since all CPUs elaborated only one function (or process) at time, given that only the main thread was available.

To this purpose, the implementation of a parallel architecture has been necessary to maximize the elaboration efficiency of shared-memory multi-core computers, and therefore exploit the entire power of their CPUs, strongly reducing computation times. To be clearer, the parallelization of a certain process consists in subdividing the main process in smaller sub-processes, called “threads”, and assigning the elaboration of a thread to a certain CPU. Hence, in this way it is possible to compute more processes (threads) at the same time, since threads are elaborated according to a parallel way, by computer CPUs. Parallelization, therefore gives the possibility of reducing computing

times, exploiting multi-cores architectures that can execute in a parallel fashion more threads.

Anyway, the implementation of this parallel architecture had not been easy for the microscopic model, because it required a partial modification of many functions in the source code, which had to be made compatible with a parallel execution, without infringing train movement rules or altering simulation outputs.

Moreover, a first step has been dedicated to measure the computing time of each code function, in order to identify the most critical one and/or the bottlenecks in terms of computational efficiency. After that these critical functions and bottlenecks had been found, a rewriting phase of their respective codes was performed, in order to convert their execution from a serial to a parallel mode.

In particular, the most time-consuming functions and bottlenecks of the whole simulation cycle, were in the process addressed to the updating of train characteristics (and in particular the function for calculating the point P at which the train had to start braking to satisfy speed-distance conditions of the objective point). Therefore, it had been necessary to entirely rewrite this part of code (which was very consistent) in a parallel fashion, and hence to sub-divide this main process in more threads. The solution at this problem has been identified, considering the “updating process relative to a single train” as a thread. In fact, in the serial version, a certain train could be updated only if the preceding trains (i.e. the trains with an earlier departure time) had already been updated. As already said before, this aspect induced a strong computing inefficiency in the simulation process. Considering instead the “updating process regarding a single train” as a thread, it is possible to update all trains on the network at the same time, since each thread is elaborated by a CPU of the computer. In this way in fact, all trains are updated together independently of their departure order, and after that their position at instant t has been calculated, both signalling and interlocking systems are consequently updated. Moreover, implementing parallelization in this way no changes or alteration is brought to the simulation process. In fact for a certain time instant t , it is not necessary that a train T2 must be updated after the preceding train T1, since this simulation model has a time-discrete nature and therefore only at instant $t+\Delta t$, T2 will know the position occupied by T1 at instant t , and only after that signal aspects have been updated too. Given that, it has been possible to parallelize train updating

process without any formal or working problem. Figure 83 shows how train updating processes were elaborated within the serial version of the model. Since a single main thread was available, CPUs could update only one train a time according to their chronological order of departure, and only after interlocking and signalling system configuration could be updated too.

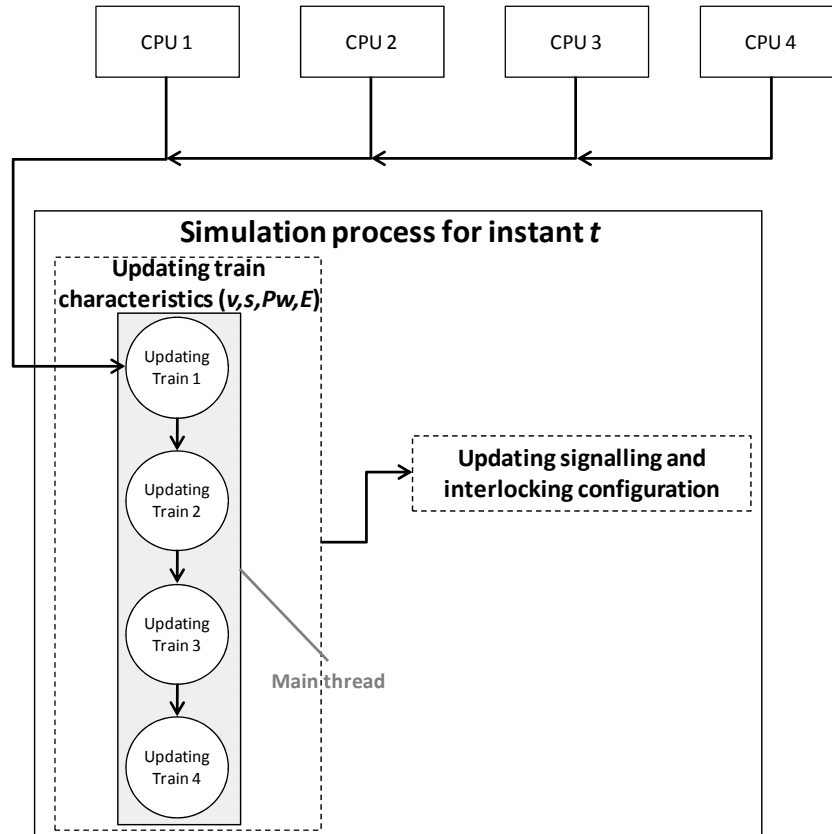


Figure 83. Serial processing of train characteristics updating: all CPUs can elaborate only a process a time since a single main thread is available.

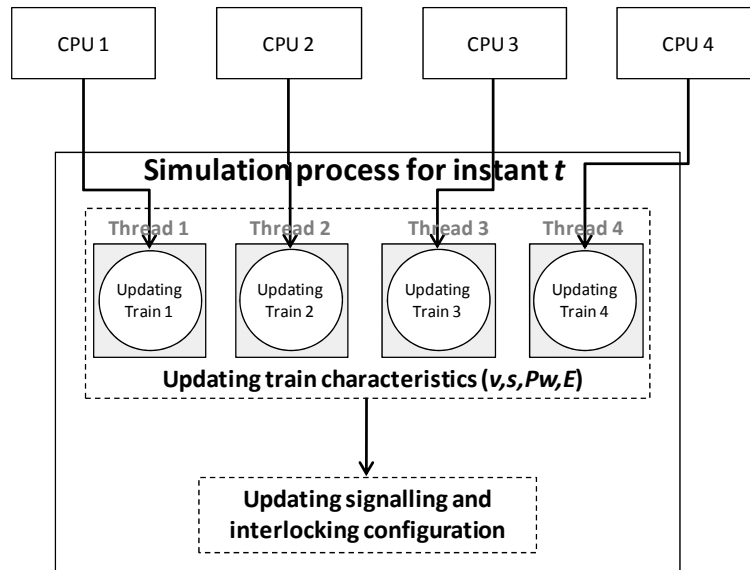


Figure 84. Parallel processing of train characteristics updating: each CPU elaborates a thread (constituted by the updating of a single train) and all trains can be therefore updated at the same time.

Figure 84, illustrates instead how trains are updated within the parallel version of the simulation model. As can be seen, the main thread is sub-divided in more threads (each one coinciding with the updating process of a single train), and each one of these (called also “slave threads”) is processed by a certain CPU. In this way, all trains can be updated contemporarily guaranteeing a reduction in computing times which can be very consistent with respect to sequential execution, especially when simulating large-sized or congested networks.

Specifically the management of threads during parallel processing has been realized through adopting the parallelization paradigm OPEN MP. OPEN MP is in fact an API (Application Programming Interface) that supports multi-platform shared memory multiprocessing programming on most processor architectures and operating systems, including Linux, and Microsoft Windows platforms, and consists of a set of compiler directives, library routines, and environment variables that influence run-time behaviour. This API has been involved in the model source code including the name of the relative header “omp.h”. Thanks to this paradigm, both the creation and the management of the so-called “slave threads” has been possible, as well as the assignment of their processing to the different computer CPUs.

Anyway one of the principal issue to treat when implementing a parallel architecture regards the number of slave threads to set within the parallel processing. This number in fact strongly influences the performances of parallelization, conditioning therefore the

execution time. Some studies, suggest that the best computing performance can be obtained when the number of threads is equal to the number of CPUs available on the computer. Actually, it depends on the number of threads that a CPU can contemporarily manage, and to establish the number for which parallelization performances are optimized is always necessary to carry out a specific test. Such test has been realized on a 4 quad-core processors 2.92 GHz shared-memory server, showing that the best computing performances are obtained when the number of slave threads is equal to the number of simulated trains.

Theory on concurrent programming, teaches that a good parallelization would consent a greater reduction of computational times, with the increasing of problem dimensions. To this purpose, three different case studies have been considered for the test, each one with a different dimension. In particular for each case study the number of slave threads has been varied, and for each different number of threads the corresponding computing time was measured. Then such times have been compared with the serial computational time to measure the so-called Speed-Up, which is simply obtainable as represented by the equation below:

$$Speed_Up = \frac{T_{serial}}{T_{parall}} \quad (36)$$

where T_{serial} is the computing time when the simulation is executed in a serial way, while T_{parall} constitutes the parallel computing time. Hence, the Speed-Up represents the number of times computing time is reduced, thanks to concurrent execution. Anyway, the first case study examined is of small dimensions and considers a railway network with a diameter of 15 km, 15 trains and 1 hour of simulation period.

The second case study, instead has medium dimensions and regards a railway network with a diameter of 26 km, 25 trains and 1 hour as simulation period.

The third case study considered has a large size since the railway network has a diameter of 100 km, with 60 trains during 2 hours of simulation period.

Simulation experiments have been carried out for each one of the aforementioned cases, varying the number of slave threads and measuring the corresponding parallel execution time. Numeric results are reported in Table 1, Table 2 and Table 3, which respectively refer to the small, the medium and the large size case studies.

T _{serial} (s)	N° Threads	T _{parall} (s)	Speed-UP
18.194	1	18.973	0.96
	2	16.204	1.12
	3	15.692	1.16
	4	14.377	1.27
	5	14.01	1.30
	6	12.737	1.43
	7	12.522	1.45
	8	11.855	1.53
	9	11.654	1.56
	10	11.473	1.59
	11	10.235	1.78
	12	10.644	1.71
	13	10.086	1.80
	14	10.11	1.80
	15	10.374	1.75
	16	10.141	1.79
	60	10.382	1.75

Table 1. Speed-Up obtained for different numbers of slave threads for the parallel processing of the small-size problem.

T _{serial} (s)	N° Threads	T _{parall} (s)	Speed-UP
58.927	1	50.6	1.16
	2	41.4	1.42
	3	32.553	1.81
	4	28.815	2.05
	5	26.554	2.22
	6	24.237	2.43
	7	23.75	2.48
	8	21.43	2.75
	9	21.078	2.80
	10	20.207	2.92
	11	18.826	3.13
	12	17.563	3.36
	13	17.95	3.28
	14	17.306	3.41
	15	17.553	3.36
	16	18.01	3.27
	17	17.786	3.31
	18	16.677	3.53
	19	16.579	3.55
	20	16.244	3.63
	21	15.15	3.89
	22	15.4	3.83
	23	15.91	3.70
	24	15.253	3.86
	25	15.189	3.88
	28	15.44	3.82
	60	15.834	3.72

Table 2. Speed-Up obtained for different numbers of slave threads for the parallel processing of the medium-size problem.

T _{serial} (s)	N° Threads	T _{parall} (s)	Speed-UP
735.418	1	753.01	0.98
	2	598.407	1.23
	3	484.859	1.52
	4	417.954	1.76
	5	322.062	2.28
	6	304.806	2.41
	7	274.483	2.68
	8	251.881	2.92
	9	245.187	3.00
	10	238.234	3.09
	11	209.247	3.51
	12	217.55	3.38
	13	196.431	3.74
	14	197.075	3.73
	15	184.347	3.99
	16	179.489	4.10
	20	159.991	4.60
	25	147.0089	5.00
	30	142.105	5.18
	35	144.393	5.09
	40	149.682	4.91
	45	147.98	4.97
	50	142.896	5.15
	55	145.138	5.07
	60	144.336	5.10

Table 3. Speed-Up obtained for different numbers of slave threads for the parallel processing of the large-size problem.

In Figure 85 these results have been represented in graphical form and compared. As can be clearly seen independently from the size of the problem the maximum computing benefit is reached when the number of slave threads is equal to the number of trains considered during simulation (15 for the small-size, 25 for the medium-size and 60 for the large-size). In fact over this number the speed-up value reaches a horizontal asymptote and therefore no further computational improvements are available.

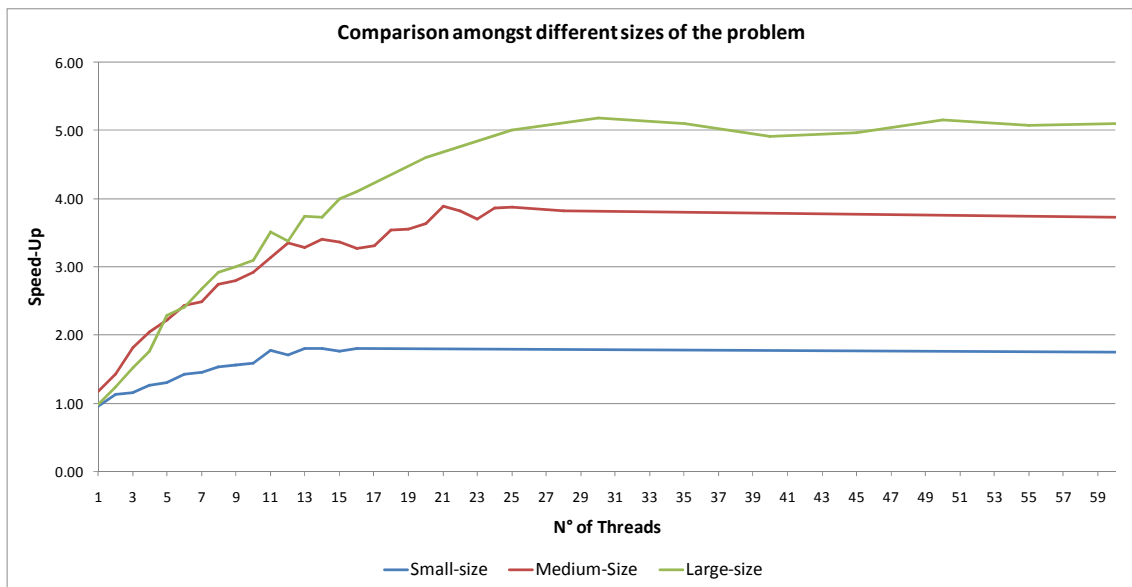


Figure 85. Speed-Up trend with respect to the size of the problem and the number of threads.

Moreover, the effectiveness of the parallelization implemented is clearly visible since the speed-up value increases with the increasing of problem dimensions. In fact for the large-size problem a reduction of more than 5 times the sequential computing times has been reached, by using this concurrent programming technique.

4.10. Model Validation

Validation of a model is one of the most important issues to take into account when considering a simulation model. In fact validation phase aims at evaluating how the simulation model is able to reproduce the behaviour of the real system. This phase is usually preceded by the so-called calibration phase, which instead can be defined as an optimization problem addressed to identify values of the input parameters of the model which minimize the distance between observed (i.e. from real system) and simulated (i.e. from the model) performance measures of the network.

Within microscopic infrastructure models, the movement of trains is usually simulated through applying mathematical laws such as the Newton's motion formula, which describes train running process as a physical phenomenon. Therefore in this case, since physical equations are involved, it is used to consider the calibration phase as a simple estimation of the values assumed in the real system by physical parameters of network components, which are taken as inputs in their corresponding models. In fact, as already explained in the previous sections, the values of input parameters which must be specified for each module coincide with those assumed by the corresponding physical parameters within the real system. Hence, for instance the values of lengths, gradients, radii of rail tracks, or weights, lengths, and characteristic curves of rolling stock will have the same values as measured in the real network.

Once that all input data have been set in the simulation model, it is of key relevance to ascertain that both the models describing the behaviour of each component and the dataset given as input within the model, are able to satisfactorily reproduce the behaviour observed in the real network. As can be easily understood, validation phase influences all results and previsions obtained by employing a simulation model, since when evaluating a certain intervention scenario, the accuracy and above all the reliability of obtained model outputs, strongly depend on how the simulation model used is "valid".

In case a simulation model is considered as not-well validated, because it is unable to describe the behaviour of the real system, it is necessary to come back to the calibration phase to identify if some error has been done in the estimation of input data, or if some phenomenon (e.g. the presence of additional resistances due to galleries) has not been taken into account within the model. Moreover in some cases it could be even necessary to intervene on the “specification” phase of the model when for example inconsistencies in the models describing the behaviour of system components do exist.

However in the specific case of the microscopic model developed, a first initial validation phase was addressed to verify if the models used to describe the behaviour of each system component, as well their interactions were able to reproduce, at least conceptually, typical dynamics of the railway system. This phase can be defined therefore as a “conceptual verification” because it has been addressed to verify if for a certain case study (which not necessarily has to be a real case study) and under certain operational conditions (e.g. under disturbed operations due to train conflicts), the model is able to return outputs which are congruent with that conditions and to consistently describe real dynamics for each component as well as for their interactions. In this way, the correctness of models specified within the “specification” phase has been investigated.

This verification has been successfully performed considering a consistent amount of different operational conditions and types of railway networks, ascertaining that outputs and reproduced dynamics were consistent with those expected. Moreover, a further verification step has regarded the comparison with a consolidated commercial microscopic railway simulation model: *OpenTrack*. In particular for several case studies, outputs returned by the realized model were compared with those given by *OpenTrack* for the same input dataset. However results of such comparisons have shown a satisfactory congruence between outputs given by the two models (for example a complete overlap between train speed-distance diagrams were observed).

Then a further validation step has been carried out, comparing results returned by the model with those observed within a real system. In particular the case study of a real mass rapid transit line has been considered: the “Cumana” line which is a metro line within the urban area of Naples city.

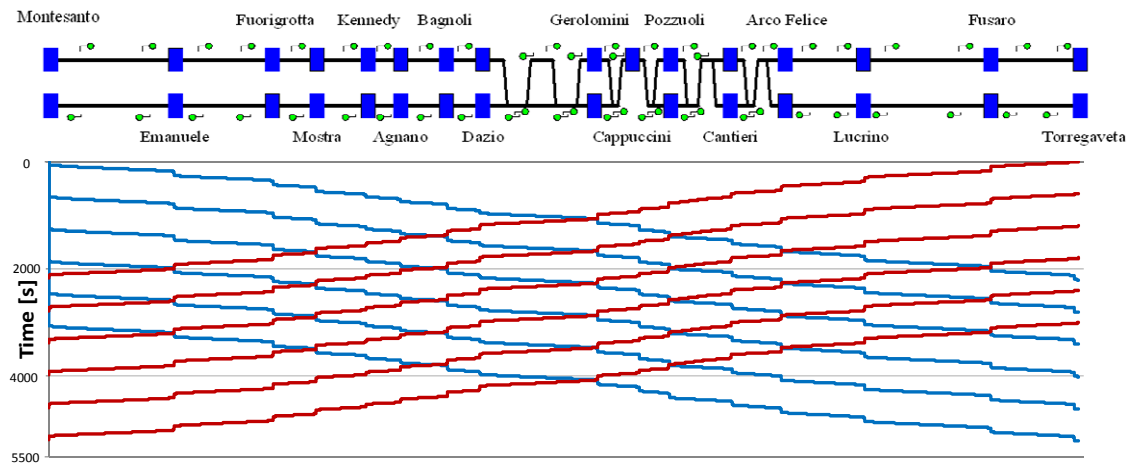


Figure 86. Schematic layout of the Cumana Line with the timetable simulated by the developed microscopic model (blue trajectories: Montesanto-Torregaveta runs, red trajectories: Torregaveta-Montesanto runs).

For this metro line all data concerning infrastructure, signalling system, as well as physical and operational train features, have been collected and then loaded as input within each module of the developed simulation model. As said before the objective of this phase is addressed to verify that simulated outputs are congruent with those observed in reality. In the specific case it has been verified that under undisturbed operational conditions, simulated train arrivals at stations, respected the real timetable.

Montesanto-Torregaveta			Torregaveta-Montesanto		
Stations	Scheduled arrival time	Simulated arrival	Stations	Scheduled arrival time	Simulated arrival
MonteSanto	08:01:00	08:01:00	Torregaveta	08:00:00	08:00:00
Emanuele	08:03:00	08:03:22	Fusaro	08:03:00	08:03:11
Fuorigrotta	08:06:00	08:06:15	Lucrino	08:07:00	08:07:14
Mostra	08:09:00	08:09:31	Arco Felice	08:10:00	08:09:44
Edenlandia	08:11:00	08:10:57	Cantieri	08:12:00	08:11:51
Agnano	08:12:00	08:12:11	Pozzuoli	08:14:00	08:13:49
Bagnoli	08:13:00	08:13:33	Cappuccini	08:16:00	08:16:00
Dazio	08:15:00	08:15:00	Gerolomini	08:18:00	08:18:09
Gerolomini	08:19:00	08:18:52	Dazio	08:22:00	08:22:15
Cappuccini	08:21:00	08:20:57	Bagnoli	08:23:00	08:23:27
Pozzuoli	08:23:00	08:23:13	Agnano	08:25:00	08:25:34
Cantieri	08:25:00	08:25:16	Edenlandia	08:26:00	08:26:40
Arco Felice	08:27:00	08:26:48	Mostra	08:28:00	08:28:31
Lucrino	08:30:00	08:29:55	Fuorigrotta	08:29:00	08:29:17
Fusaro	08:34:00	08:34:12	Emanuele	08:32:00	08:32:14
Torregaveta	08:36:00	08:36:08	MonteSanto	08:35:00	08:35:10

Table 4. Comparison between scheduled and simulated train arrivals at stations, for a single run for each direction.

Hence, the validity of the model has been tested measuring its capability in reproducing train behaviour under scheduled operations. In fact, Table 4 shows the comparison between scheduled and simulated arrivals for two trains running respectively along “Montesanto-Torregaveta” and the opposite direction. As can be clearly seen,

differences are very slight and do not exceed the amount of 40 seconds (reached at “Edenlandia” station along “Torregaveta-Montesanto” direction) for both directions. These results confirm therefore the validity of the developed model and its ability in reproducing real train behaviour under ordinary service conditions. Moreover, Figure 86 illustrates the schematic layout of the Cumana line and the graphical timetable simulated by the realized microscopic model.

Furthermore, also a validation aiming at verifying the capability of the model in reproducing system behaviour under disturbed conditions, has been realized. In particular the total train arrival delay at “Torregaveta” station has been measured simulating a period of 1 hour within a disturbed scenario, where the first train had a departure delay of 6 minutes due to a conflict with an opposite train at “Arco Felice” station. The value of the total train arrival delay returned as output by the simulation model is 12 minutes and 54 seconds which is congruent with that detected on the real system in the same disturbed scenario of 13 minutes.

4.11. Uncertainty analysis of the model: Sensitivity Analysis.

The assessment of uncertainties in model outputs is an essential phase to understand the reliability of the model itself and therefore its usefulness in supporting decisional phases relative to planning or designing activities. Such uncertainties are due to different sources which are often mixed in a complex way, and that can be attributed in part to the (in)adequacy of the models with respect to the reality, and in part to (uncertain) model inputs. In particular uncertainties due to the model inadequacy depends on different sources, such as the type of modelling assumptions, the kind of equations used, the level of space-time discretization, etc. These types of uncertainties can be effectively reduced by “improving” the model intervening on one or more of these issues. However, this usually brings the model to higher computing times (especially when the model is enriched in details), therefore the choice of the most appropriate modelling structure depends on the context in which it must be employed (e.g. off-line or on-line applications), and above all it must derive from a satisfactory trade-off between quality of results and computing times.

For what concerns instead uncertainties related to model inputs, it is possible to distinguish between “observable” and “unobservable” inputs. Specifically the former are definable as all those model inputs which have a correspondent in the real world,

and therefore whose value can be measured by direct investigations. For example in the railway field this is the case of the characteristics of railway infrastructures (e.g. radii, gradients, lengths, maximum line speed, etc.), the features of rail vehicles (e.g. number of wagons, vehicle mass, “tractive effort-speed” curve, etc.), as well as the characteristics of the signalling system (e.g. block section lengths, number of transponders, signalling delay time, etc.). The latter are considered instead as those inputs which are difficult to measure in the reality (e.g. adhesion coefficient between wheel rim and rail, average arrival/departure delay at stations), or that not have a direct equivalent in the reality (e.g. coefficients of resistances equations). These inputs are considered uncertain in order to cover both the epistemic uncertainty (which is reducible by enlarging the number of measurements of the input and/or increasing available data) and the aleatory uncertainty (which cannot be reduced since represents the stochasticity inborn in the own nature of the input itself), and therefore they can be only indirectly estimated by inverse analyses or calibration.

Obviously uncertainties in both the models employed, and the inputs propagate into model outputs, influencing the reliability of results and the usefulness of the model itself. Hence, it is immediate to understand the importance of an uncertainty analysis of the model. In particular Figure 87 depicts a conceptual framework for uncertainty estimation and management (*de Rocquigny et al. 2008*).

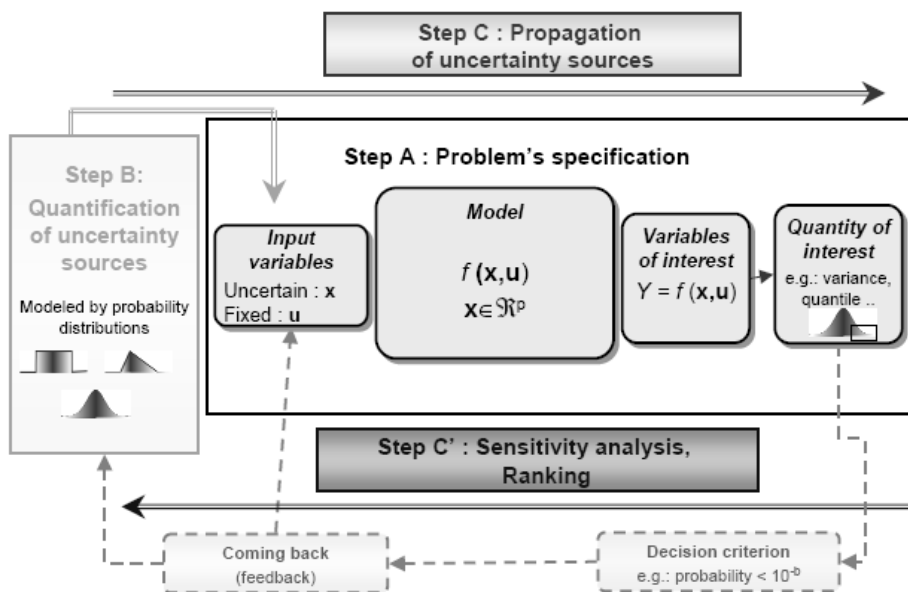


Figure 87. Conceptual framework for uncertainty assessment and management (*de Rocquigny et al. 2008*)

As can be seen Step A concerns problem's specification, which consists in defining input and output variables, specifying models and identifying the quantities of interest to measure uncertainties in the outputs. Model inputs can be constituted by both random (\mathbf{x}) and deterministic variables (\mathbf{u}), according to the choice of the analyst. In fact depending on the problem setting the random inputs \mathbf{x} may include all the sources of uncertainty such as the parametric or the model uncertainty. Other inputs instead, can be considered as deterministic (or fixed), \mathbf{u} , when they actually assume a constant value in the problem or when their variability (uncertainty) is negligible with respect to the output variables of interest.

Step B, regards the quantification of uncertainty sources, which implies the definition of the joint pdf of the uncertain inputs or their marginal pdf with simplified correlation structures or when considered as independent variables. This is usually the most expensive step of the uncertainty analysis, since it requires the collection of a large amount of information by means of direct observations (for observable inputs), expert judgments, physical arguments or indirect estimation (for unobservable inputs).

Step C is relative to the propagation of uncertainties in the model outputs, and it is therefore essential to understand how uncertainties in the models as well as in the inputs influence uncertainties in the outputs themselves. To this purpose a Monte-Carlo simulation framework is usually used. However, this phase mainly consists in estimating the pdf of the output variables of interest \mathbf{Y} , given the pdf of the random inputs \mathbf{x} , the values of the fixed inputs \mathbf{u} , and of course the model $f(\mathbf{x}, \mathbf{u})$.

Step C' corresponds instead to the so-called *sensitivity analysis* or *importance ranking*. This phase constitutes the feedback process in the complex of the uncertainty management, and consents to understand how uncertainties in the outputs are apportioned to different sources of uncertainties in the inputs. The objective of the sensitivity analysis is therefore to instruct the modeller with respect to the importance of the uncertain inputs in the determining the variables of interest. To this aim, this analysis needs of some statistical treatment of the input/output relations drawn within the uncertainty propagation step.

In particular a sensitivity analysis has been conducted in this thesis work, considering a real MRT line: the "Cumana" line of Naples (described in the previous section and represented in Figure 86). In fact, this analysis has led to understand how the variability

in a certain network performance (which is an output of the microscopic model) is partitioned to the variability of the different design variables considered (which are instead inputs of the model). Moreover as a consequence, it has been possible to identify for a certain performance measure, the design variables which mostly affect it. These results have shown therefore the usefulness of sensitivity analysis also for supporting design activities, since it sheds light on the design parameters which actually need to be modified to efficiently improve a certain network performance. Furthermore this aspect, consents to efficiently allocate economic resources by focusing the efforts only on the most relevant variables, without wasting money for intervening also on “non-key” parameters. In this sense therefore the employment of a sensitivity analysis can lead to optimize the allocation of available resources, which is an issue of key relevance today, given the increasingly reduction of project budgets.

In the following after a brief description of the “variance-based” method employed to perform the sensitivity analysis, the application to the Cumana line is illustrated and relative results are then showed.

4.11.1. Sobol’ variance-based method for performing sensitivity analysis

Many methods and techniques are available to perform a sensitivity analysis, in fact it is possible to distinguish amongst: i) Input/output scatter-plots, ii) Sigma-normalized derivatives, iii) Standardized regression coefficient, iv) Elementary effects, v) Variance-based techniques, vi) Monte Carlo filtering and vii) meta-modelling. As said before, the method used is a variance-based method, and in particular the one given by Sobol’ and then improved by Saltelli (*Saltelli et al. 2008*). This is a global method, i.e. it consents to investigate homogeneously the whole domain of the input parameters, and allows to estimate sensitivity indices with a number of model simulations that is smaller than the one required by the other methods. Such a method will be described in the following, while the illustration of the other techniques can be found in (*Saltelli et al. 2008*).

In particular a first application of variance-based method for sensitivity analysis had been observed in (*Cukier et al. 1973*) and successively generalized by Sobol to provide a Monte Carlo-based implementation of the concept. The system under investigation can be mathematically represented as:

$$Y = f(Z_1, Z_2, \dots, Z_r), \quad (37)$$

where Z_i , ($i = 1 \dots r$) represent system inputs while Y is its output (or one of its outputs if they are more than one). It is necessary therefore, to understand the implications for the variability of output Y when a certain input variable Z_i is set to a specific value z_i^* . The resulting variance of Y , called as conditional variance is:

$V_{Z \sim i}(Y | Z_i = z_i^*)$, indicating with the symbolism $Z \sim i$ that the variance across all the variables has been estimated, but the i^{th} . It is expected that the conditional variance will be as lower than the total variance of Y as bigger is the influence of the variable Z_i . Hence, the conditional variance can be considered as an index of the sensitivity for Z_i . However, since this sensitivity index is calculated taking into account only the value z_i^* of the input parameter, it is necessary to refer to the average of such index over all possible points z_i^* and therefore $E_{Z_i}(V_{Z \sim i}(Y | Z_i = z_i^*))$.

In particular the unconditional variance of the output Y , can be expressed as:

$$V(Y) = E_{Z_i}(V_{Z \sim i}(Y | Z_i)) + V_{Z_i}(E_{Z \sim i}(Y | Z_i)). \quad (38)$$

Therefore it is clear that if parameter Z_i is significant, then the value of $E_{Z_i}(V_{Z \sim i}(Y | Z_i))$ will be small, thus the closer is $V_{Z_i}(E_{Z \sim i}(Y | Z_i))$ to $V(Y)$, the higher will be the influence of Z_i on the output Y . For this reason it is possible to define the so called first order sensitivity index (Saltelli *et al.* 2008) of Z_i with respect to Y :

$$S_i = \frac{V_{Z_i}(E_{Z \sim i}(Y | Z_i))}{V(Y)} \quad (39)$$

First order sensitivity index is a very important measure to understand how much the variability of a certain output is influenced by the only input parameter Z_i . Furthermore coupling equations (38) and (39) it is immediate to write that $S_i \leq 1$. Anyway a certain model can be defined as additive when:

$$\sum_{i=1}^r S_i = 1 \quad (40)$$

Actually in this case, the unconditional variance of the model can be decomposed in the sum of the first order effect of each single variable. Usually this is not the case, meaning that the joint combination of some variables can be responsible for a certain share of the unconditional variance, that is just the definition of non-additive models. In this case, a

low first order sensitivity index does not necessarily imply that the corresponding variable has a scarce effect on the output variance, since it might considerably contribute to the total output variance, by means of its combination with the other variables. For this reason, using the so-called ANOVA-HDMR (Analysis of Variance-High Dimensional Model Representation) decomposition developed by Sobol (*Sobol* 1993), it is possible to say that a full analysis of a model with r input variables requires for all the elements of the following equation to be discovered (in number of $(2^r - 1)$):

$$\sum_{i=1}^r S_i + \sum_{i=1}^r \sum_{j>1}^r S_{i,j} + \sum_{i=1}^r \sum_{j>l>j}^r S_{i,j,l} + \dots + S_{1,2,3,\dots,r} = 1. \quad (41)$$

The computation of all the sensitivity indices in equation (41) would require a very expensive experimental work. To reduce the efforts another indicator can be defined and coupled with the first order sensitivity index: the total effects index defined as (*Homma and Saltelli* 1996, *Saltelli* 2002):

$$S_{T_i} = 1 - \frac{V_{Z_{\sim i}}(E_{Z_i}(Y | Z_{\sim i}))}{V(Y)} = \frac{E_{Z_{\sim i}}(V_{Z_i}(Y | Z_{\sim i}))}{V(Y)}. \quad (42)$$

Such index provides in fact for the input parameter Z_i , the sum of all the elements in equation (41), also taking into account the variance due to the i^{th} factor itself. When the total index is equal to 0, the i^{th} parameter can be fixed without affecting the outputs' variance. If instead $S_{T_i} \approx 0$, the approximation made depends on the value of S_{T_i} (*Sobol et al.* 2007). It is worth noting that while $\sum_{i=1}^r S_i \leq 1$, $\sum_{i=1}^r S_{T_i} \geq 1$, both being equal to 1, only for additive models.

However, the evaluation of both the first-order and the total sensitivity indices can be realized following the procedure described in (*Saltelli et al.* 2010). In particular defining N as the size of the Monte Carlo experiment, it is necessary to generate two (N, r) matrices of quasi-random numbers drawn according to the Sobol sequence (*Sobol* 1976). Using these two matrices of quasi-random numbers, two corresponding matrices of values for the input parameters (contained within the domain of each one) of the model reported in (37) are generated. Such matrices of input values can be respectively named A and B , and represented as:

$$A = \begin{bmatrix} z_1^{(1)} & z_2^{(1)} & \dots & z_r^{(1)} \\ z_1^{(2)} & z_2^{(2)} & \dots & z_r^{(2)} \\ \dots & \dots & \dots & \dots \\ z_1^{(N)} & z_2^{(N)} & \dots & z_r^{(N)} \end{bmatrix} \quad (43)$$

$$B = \begin{bmatrix} z_{r+1}^{(1)} & z_{r+2}^{(1)} & \dots & z_{2r}^{(1)} \\ z_{r+1}^{(2)} & z_{r+2}^{(2)} & \dots & z_{2r}^{(2)} \\ \dots & \dots & \dots & \dots \\ z_{r+1}^{(N)} & z_{r+2}^{(N)} & \dots & z_{2r}^{(N)} \end{bmatrix} \quad (44)$$

Then a set of r matrices A_B is obtained assembling r matrices equal to A except for the i^{th} column (with i varying from 1 to r among the r matrices) that is taken from B . Hence:

$$A_{B,i} = \begin{bmatrix} z_1^{(1)} & z_2^{(1)} & \dots & z_{r+i}^{(1)} & \dots & z_r^{(1)} \\ z_1^{(2)} & z_2^{(2)} & \dots & z_{r+i}^{(2)} & \dots & z_r^{(2)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ z_1^{(N)} & z_2^{(N)} & \dots & z_{r+i}^{(N)} & \dots & z_r^{(N)} \end{bmatrix} \quad \text{for } i = 1 \dots r \quad (45)$$

At this point, the model is evaluated for all the $[N \cdot (r+2)]$ combinations of input variables as given by matrices A , B and A_B so as to produce the vectors of outputs $y_A = f(A)$, $y_B = f(B)$, $y_{AB,i} = f(A_{B,i})$, for $i = 1 \dots r$.

Such vectors are sufficient for the evaluation of all the first-order and total effects indices. Moreover the sensitivity indices can be evaluated using the estimator given by Jansen (*Jansen 1999, Jansen et al. 1994*) and reported in (*Saltelli et al. 2010*):

$$S_i = 1 - \frac{\frac{1}{2N} \sum_{j=1}^N (y_A^{(j)} - y_{A_{B,i}}^{(j)})^2}{V(Y)}, \quad (46)$$

$$S_{T_i} = \frac{\frac{1}{2N} \sum_{j=1}^N (y_A^{(j)} - y_{A_{B,i}}^{(j)})^2}{V(Y)}. \quad (47)$$

Furthermore it is necessary to say that no universal recipe exists for determining the size of the Monte Carlo experiment N . This dimension can in fact vary from few hundred to several thousands. A possible strategy to be adopted is to choose an initial value of N and verify that for this number, the plots of the sensitivity indices have reached a stable value (i.e. a value for which they do not depend anymore from N). In case they are not stable, it is necessary to adopt a larger value of N . For this reason, in the following, the results of the sensitivity analysis (i.e. sensitivity indices) will be plotted against N to ascertain that they reach a stable value.

However, the procedure described above, can be applied to models where the input parameters are not correlated. But in the case of correlated inputs it is necessary to adopt different strategies to solve this problem, as reported in (*Saltelli and Tarantola 2002, Jacques et al. 2006*). In the case of a railway system for example input parameters as service headway and block section length are correlated, since the headway between two consecutive train runs cannot be lower than the signal headway calculated for a certain block section length and a determined signalling system (for reasons regarding safety and blocking time conflicts, see *Hansen et al. 2008* for reference). To solve this issue, all the output values $y^{(j)}$ (being a general value of y_A, y_B or y_{AB}) corresponding to those rows $z^{(j)}$ of the input values matrices (A, B or A_B), where the drawn value of train headway is lower than the signal headway relative to the drawn value of block length, have been discarded from the analysis.

4.11.2. The case of a MRT system: the Cumana line.

The Sobol' "variance-based" method has been applied for performing a sensitivity analysis for the Cumana line, whose schematic layout has been illustrated in the previous section (see Figure 86). In particular this line is currently equipped with an old signalling system dating back to the 60's, which is partially constituted of track circuits while train movements are mainly regulated by relay interlocking systems located within each station area and operated via distant control by a centralized traffic control centre positioned in "Montesanto" station. However such signalling system have been modelled during simulation as a track circuit multi-aspect system, and such assumption had been also successfully confirmed by previous simulation experiments performed for this line. The minimum line headway is 450 seconds, and the maximum line speed limit is set to 90 Km/h. The train headway, fixed by the current timetable is 10 minutes, while

the scheduled dwell time is set to 30 seconds for all stations. To correctly simulate train performances, overall in terms of punctuality, a preliminary phase consisting in the determination of train dwell time distribution at stations had been necessary. In particular real station dwell time data were collected for each station of the line and then aggregated to find which distribution function matched best with data and was more representative for train stopping operations at stations. Different distributions were tested but the one which showed the highest p-value of the Kolgomorov-Smirnov test was the log-normal distribution, with a mean value $\mu = 3.29$ s and a standard deviation $\sigma = 0.399$ s. Figure 88 shows the outcomes of this study for the specific case of “Fuorigrotta” station. As can be seen log-normal distribution (red line) actually fits well with the observed distribution of station dwell times (histograms).

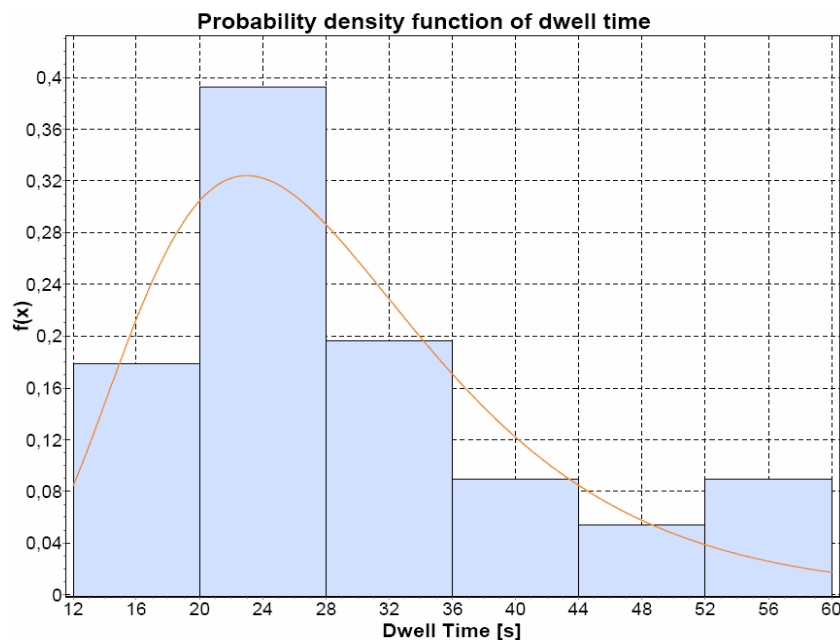


Figure 88. Cumana line: log-normal fit of dwell time distribution at “Fuorigrotta” station.

Dwell times at stations have been modelled within simulation, as independently and identically distributed (i.e. stations are not correlated and the same distribution parameters have been considered for all stations) according to a log-normal random variable. Such assumptions are moreover in line with results presented in literature by previous works on statistical dwell time modelling for metro systems (*Martinez et al.* 2007).

Anyway, such mass transit line is currently under a renewal phase where network operators need to determine effective infrastructural and/or operational interventions in

order to improve overall system performances with particular reference to punctuality and energy consumption rates of train runs. Furthermore, given the limited financial budget available for investment, it is strongly recommended to optimize economic resources intervening on those design variables which mostly influence the considered network performances.

In particular 4 different measures of performance Y have been considered within the analysis:

- *Average train arrival delay (Av_Delay)* at final station (in seconds), calculated as the ratio between the total train arrival delay at final station (i.e. the sum of arrival delays with respect to scheduled arrivals over all train runs) and the number of simulated train runs. Obviously train delays will be measured at “Torregaveta” station for trains running along “Montesanto-Torregaveta” direction, and at “Montesanto” station for trains running along the opposite direction. In particular the convention used here is the same reported in (Yuan 2006) and consists in considering the values of train delays as negative when trains arrive ahead of schedule, and positive while delayed arrivals are performed, For this reason such performance measure can assume both negative and positive values.
- *Standard deviation of train arrival delay (Std_Delay)* at final station (in seconds), to understand which parameter has a greater influence on the variation of train arrival delays. Such measure of performance has been calculated through adopting the traditional formula of the standard deviation:

$$Std_Delay = \sqrt{\frac{\sum_{t=1}^T (Train_Delay_t - Av_Delay)^2}{T - 1}}, \quad (48)$$

where $Train_Delay_t$ is the arrival delay relative to train t , while T is the number of simulated trains.

- *Average train energy consumption (Av_ECons)*. This measure has been expressed in Mega Joule (MJ) and considered as the ratio between the total energy consumption (i.e. the sum of energy consumption over all train runs) and the total number of simulated trains.

- *Standard deviation of energy consumption (Std_ECons)*. Also expressed in Mega Joule (MJ), this measure is necessary to comprehend which parameters mostly condition the variation of train energy consumed during service. Also for this one the classical standard deviation formula has been employed:

$$Std_ECons = \sqrt{\frac{\sum_{t=1}^T (Train_ECons_t - Av_ECons)^2}{T-1}}, \quad (49)$$

where $Train_ECons_t$ represents the energy consumed by train t , while T is the number of simulated trains.

The analysis has been conducted considering the four following input parameters Z_i ($i = 1...4$) :

- *Block section length (BSL)*, considering that the line is equipped with an equi-block three-aspect signalling layout. The variability range considered for this parameter is $\{0.200, 1.500\}$ Km, where in particular the lower bound is the minimum section length which assures a safe braking with a three-aspect system and for a maximum speed of 90 km/h (see *Gill and Goodmann* 1992 for reference on the calculation methodology).
- *Train headway (HW)*, whose domain has been set to $\{120, 720\}$ s, where in particular the lower bound represents the minimum line headway (signal headway) corresponding to the lower bound of block section length (0.200 Km) with the considered signalling system. As already said, this parameter is correlated with the block section length, since headways must be higher than the signal headway to guarantee a safe and conflict-free scheduled service. Therefore simulation experiments relative to all those rows of the input values matrices (A , B and A_B) for which the drawn headway value was lower than the signal headway corresponding to the drawn block section length, have been discarded from the analysis.
- *Average (Av_Dwell) and Standard Deviation (Std_Dwell)* of station dwell time distribution function, supposing as said before that dwell times are identically and independently distributed as a log-normal random variable for all stations. Therefore

such parameters tallies with the parameters of the normal distribution according to which logarithms of dwell times are distributed (by definition of log-normal distribution). In particular, for the logarithms of dwell times the range assumed for the average is $\{2.99, 3.41\}$ s while for standard deviation is $\{0.125, 0.7\}$ s.

Results

As previously illustrated within the previous section, the total number of simulation model evaluations necessary to perform the described sensitivity analysis coincides with the number of all possible combinations of the input variables, and therefore depends on the number of these latter which here is $r = 4$, as well as on the size of the Monte Carlo experiment. In particular, this latter has been set to $N = 10000$, since a preliminary study confirmed that for this number stable values of the sensitivity indices are obtained. Therefore a total of $N \cdot (r+2) = 60000$ simulation model estimations must be carried out. Actually results of each one of these model estimations are here considered as the average of results from 10 different simulation replications (using different seeds in the random number generation process of the simulation), in order to reduce the impact of stochasticity with a confidence of 90%, following the indications reported in (Law 2007). Hence a total of 600000 simulation runs must be performed. Indeed, due to the aforementioned correlation between train headway and block section length, a certain amount of model evaluations have been discarded from the analysis and only 2000 simulation experiments have been really usable for performing the analysis (hence $N \cdot (r+2) = 12000$ simulation model estimations and 120000 simulation runs). Results obtained considering a simulation period of 1 hour, are illustrated below.

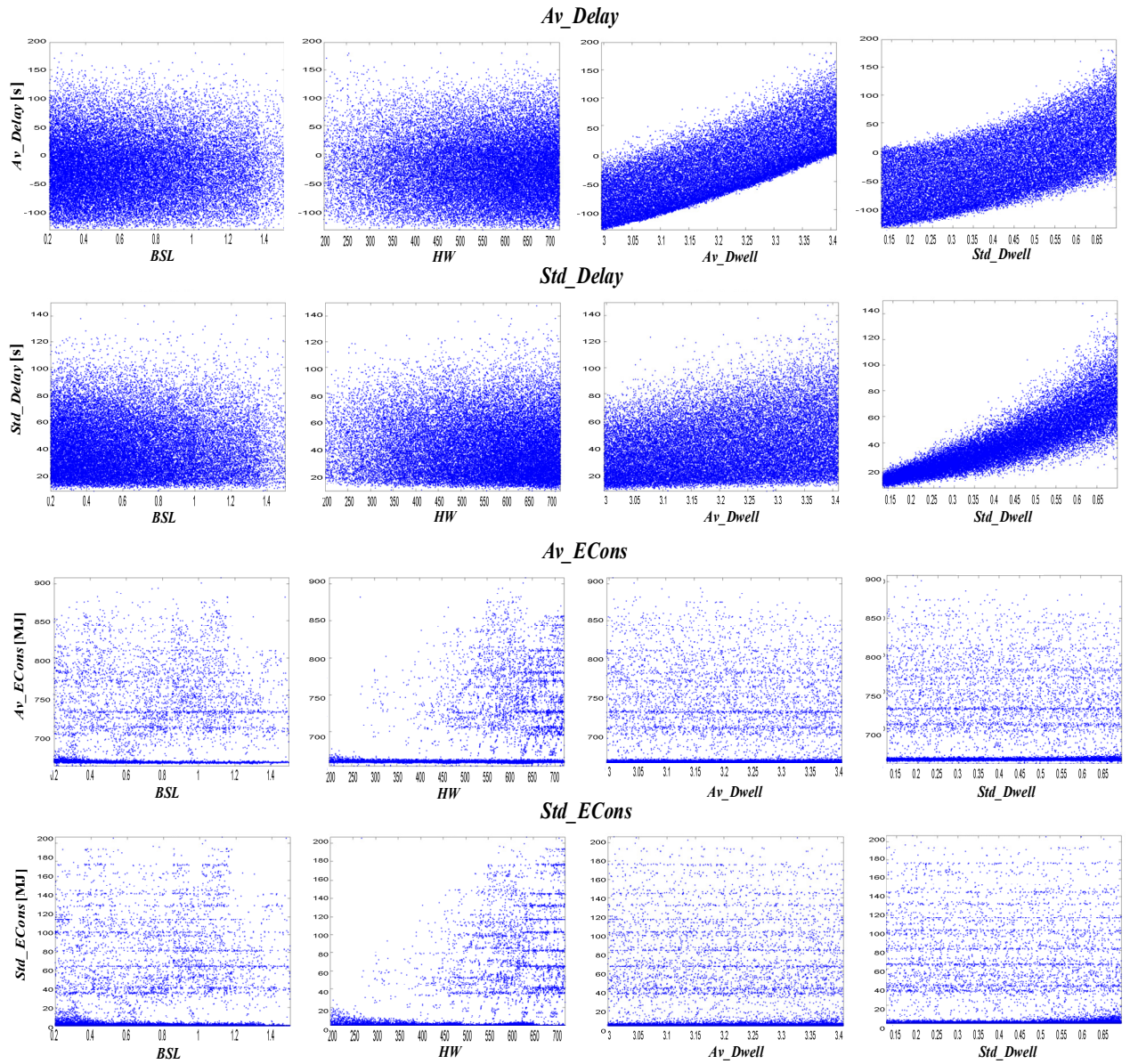


Figure 89. Scatter-plots of each considered output against each one of the input parameters

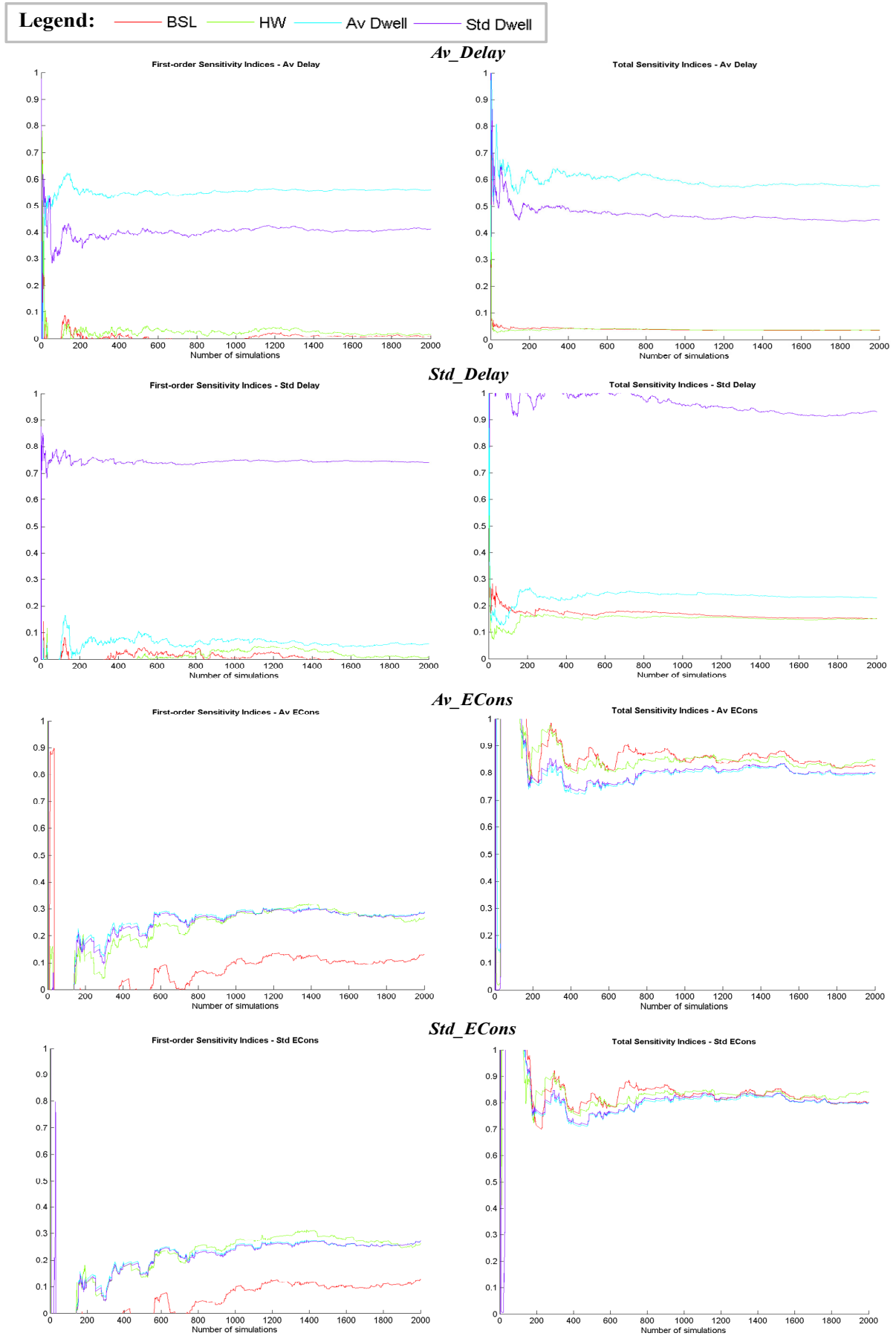


Figure 90. First-order (left) and Total (right) sensitivity indices of each input for the 4 outputs.

In particular Figure 89 shows for each output the corresponding scatter plots against each one of the considered input parameters.

As can be seen, these plots immediately underline for the average train arrival delay Av_Delay , that it seems to be more influenced by dwell time distribution parameters Av_Dwell and Std_Dwell , as for increasing values of these parameters, it is possible to observe an increasing trend of Av_Delay . The standard deviation of arrival delays Std_Delay , seems instead to be mostly conditioned by the standard deviation of dwell time distribution, since it is clearly visible that this output tend to increase when Std_Dwell increases. For what concerns both the average and the standard deviation of train energy consumption, nothing can be said with clarity by their plots and the estimation of the sensitivity indices become fundamental to this purpose. Moreover, from their plots against both train headway and block section length, the effect of the correlation between these two input parameters is highlighted, since for low headway values as well as high block section lengths, the population of plotted points is less crowded (in fact the higher is the block section length, the higher will be the signal headway and therefore the lower will be the probability that a low headway value will be compatible with that length).

Instead in Figure 90, for each performance measure the relative graphs of both the first-order and the total sensitivity indices have been reported for each one of the considered inputs.

As illustrated in the previous section, an input parameter will be more influent for a certain output when the higher are the values of its first-order and total sensitivity indices. As can be seen, for the performance measure Av_Delay , both the first-order and the total indices diagrams, confirm what appeared from the previous scatter plots analysis, i.e. that the most influent parameters are constituted by both the average and the standard deviation of station dwell times. The same observation can be made for the standard deviation of arrival delays Std_Delay , since as already emerged from the relative scatter-plots, also first-order and total indices graphs clearly show that Std_Dwell is the input which mostly affect this output. These results therefore underline for the analyzed MRT line, that platform phenomena (e.g. platform congestion) at stations mostly condition the variability of train arrival delays, and therefore their punctuality. This means that the total arrival delay of train runs, (calculated as the sum

of both original and knock-on delays) is mostly due to original delays experienced at stations rather than to knock-on delays (whose effects could be instead reduced by acting on headways and block section lengths).

Certainly, the design of a robust timetable, optimizing the allocation of both recovery supplements to train runs as well as buffer times between consecutive runs, is a task which must be always performed to guarantee a stable service. Robust timetabling, can be in fact considered as a “passive defence” solution against the propagation of train delays on the network, in the sense that it does not act directly on the causes of delays (i.e. on original delays) but mainly mitigates their effects. It is sure that original train delays cannot be unfortunately eliminated from the system, since their stochastic nature, but when infrastructural interventions are under examination as in the case study here presented, something could be done to try to reduce their causes. In this case results of the sensitivity analysis carried out, clearly highlight for *Av_Delay* and *Std_Delay* the influence of both the two dwell time distribution parameters. It is clear that not so many things can be done to reduce *Av_Dwell*, since this parameter mainly depends on the flow of alighting and boarding passengers at stations (*Qi et al.* 2008), which is certainly not controllable. Some actions could be instead realized for reducing *Std_Delay*, which mainly depends on fluctuations of dwell times due to matters like: platform congestion phenomena, presence of people with mobility difficulties (e.g. old people), etc. In fact “active defence” interventions (which directly act on the causes) could be applied to this purpose, for example acting on the layout of station platforms (enlarging their dimensions, or improving their configuration with respect to passengers mobility), equipping platform areas with ATO (Automatic Train Operation) systems like station stop beacons (and door enabling loops) which assure the train stop safely at the correct position towards the platform (therefore also reducing disturbances due to incorrect train positioning), or intervening on the rolling stock, buying for instance vehicles with larger passenger doors (or with more doors for coach) as well as equipped with a low floor in order to ease alighting and boarding operations of passengers also within congested conditions.

As regards instead the average and the standard deviation of train energy consumption, it is possible to see from the diagrams of total sensitivity indices, that all the input parameters have a certain influence on these outputs but above all the block section length and the headway. It is worth noting that while *BSL*, has the lower influence with

respect to the other parameters, when taken by itself (since has the lower values of the first-order sensitivity index), it becomes the most influent parameter when its interactions with the other inputs are considered (see the values of the total sensitivity index). Moreover, unlike arrival delays, the obtained results highlight that the variability of train energy consumption, is mainly due to knock-on disturbances rather than original disturbances. In fact, when a train follows a delayed train, it may be more probable that it has to perform more unscheduled decelerations and accelerations due to some restricted aspect of signals, and therefore that it consumes more energy (because of the higher number of acceleration phases and the larger running times) than undisturbed running. Furthermore as confirmed by the total index estimation (and in accordance with literature), it is necessary to intervene on train headway and block section layout to prevent strong knock-on disturbances and therefore higher energy consumptions. In this case the actions on both these parameters can be seen as an active defence solution since they directly touch the cause of the problem, i.e. the propagation of disturbances on the network. Hence, possible solutions to reach this purpose could be: acting on the timetable, enlarging train headways and therefore buffer times between consecutive trains, or intervening on the signalling system for example reducing the length of block sections (and reduce the propagation of running disturbances) or changing its technology (upgrading for instance from a track circuit system to an ETCS level 1).

Moreover, what must be noted is that results of a sensitivity analysis are fundamental to understand also for a complex problem which are the real causes and therefore allow to ease its resolution by focusing the attention only on the most sensitive variables. This is one of the reasons why a preliminary sensitivity analysis should always be realized particularly when facing complex problems like the design of railway systems.

Chapter 5. Practical applications of the microscopic model to support different design activities.

5.1. Introduction

In this chapter some applications are illustrated to show potentialities of the developed microscopic model and above all, to better understand which kind of analysis and in which phase it can be employed to support design activities. Moreover, the implementation of such applications has been necessary to further test the validity of the model itself and obviously its suitability in being interfaced with external applications (e.g. via API) or mathematical structures such as optimization models and probabilistic analyses frameworks. In particular, different kinds of analyses have been carried out, concerning with different problems that can be considered during the wide range of activities relative to the design process of a railway system. Specifically, a first application has regarded the evaluation of different intervention scenarios on both signalling system layout and service headway, for a real mass rapid transit case study, employing a classic “what-if” approach. Then a second application has consisted in integrating the microscopic model within a “black-box” optimization loop to design an optimal signalling system layout in order to assure a required level of system capacity (therefore a required minimum line headway) and contemporarily minimize investment costs (in terms of the number of block sections and signalling circuits). Therefore in this case a “what-to” design approach has been used. Successively this “black-box” optimization loop has been employed also in finding the equi-block signalling layout which guarantees the best trade-off between user’s generalized cost and investment costs. Furthermore, other applications have been implemented in the field of the so-called RAM analysis, where for a certain failure scenario different recovery strategies have been evaluated to establish the one which consented to satisfy the RAM indexes established by contract requirements (for example in terms of train punctuality) as well as a determined level of Quality of Service delivered to passengers. Since the RAM analysis is a probabilistic analysis addressed to make inference on failures of network components, millions Monte-Carlo simulations of the model are needed to evaluate the effects on system performances when a stochastic failure event is drawn. However in this case, the computation of this large amount of model evaluations would require unreasonable computing times if a microscopic model was used, since as said before

this kind of model is inefficient for performing probabilistic analysis. Therefore to solve such problem a dynamic integration with an efficient “own-built” mesoscopic model based on the SAN (Stochastic Activity Networks) formalism is under consideration, in order to realize Monte-Carlo simulations of ordinary service exploiting the computing efficiency of the mesoscopic model and activating the microscopic model only when a failure event is drawn to accurately estimate its effects on the network. Moreover, such integration would lay the groundwork for the development of a complete simulation framework of railway systems which takes into account also interactions with the demand-side, interfacing a module for simulating the dynamics of supply side components (e.g. using the proposed hybrid mesoscopic-microscopic model) with a module for simulating effects on passenger demand flows.

In the following of this chapter, each one of the aforementioned applications will be deeply described.

5.2. Design of signalling system using a “what-if” approach for evaluating intervention scenarios

As known, signalling system strongly influences the capacity of railway networks since it controls and safely regulates train movements on the track. Due to the recent increase of demand for passengers and freight transportation on railway networks, train operators and above all infrastructure managers, need to adequate system capacity to the required traffic levels and therefore must identify effective intervention solutions which are able to meet such objectives. To this purpose are more and more increasing throughout the world works aiming at the realization of new railway networks equipped with high-capacity signalling systems (e.g. ETCS level 1 or level 2), or “re-signalling” works addressed to entirely substitute the existing signalling layout to increase capacity levels. However, designing phases which aim at the determination of the most effective signalling system layout, need to be supported by a microscopic simulation model to accurately evaluate the effects induced by the different intervention scenarios. In particular, as already shown within Chapter 2, once the signalling system layout has been established, the corresponding value of capacity can be calculated simulating train runs, determining their respective “blocking time” stairways (for each block section and according to the working features of the signalling type), and identifying the critical section which dictates the minimum headway that can be safely performed for that line (the so-called “signal headway”). To this purpose, an apposite module for the

calculation of the “signal headway” corresponding to a certain signalling system layout has been developed in C++ and added to the source code of the microscopic model. Practically, such a module simply applies the “blocking time theory” to estimate the signal headway of the system. Therefore it first performs deterministic simulations of train runs, then calculates their minimum running times on the line, and according to the features of the signalling system implemented (e.g. communication delay time, approach time, release time), determines for each block section the corresponding “blocking time”, identifies the critical section and computes the signal headway.

In particular, a mass rapid transit system of the urban area of Naples has been considered: the Cumana line (whose scheme has been illustrated in Figure 86). Recently, the infrastructure manager of this line needed to increase the capacity level of the network to face a foreseen increase of passengers demand. In this application it has been supposed that they need to decrease the signal headway of the line (i.e. increase its capacity) in order to assure a service headway of five minutes. Since the current signalling layout of such system (based on an old centralized electrical interlocking system which supervises the traffic within each station area) does not allow to meet the capacity level required, opportune infrastructural and “re-signalling” works have been taken under consideration. In particular, two different intervention scenarios have been envisaged:

- 1) Intervening only on the infrastructure layout realizing a line entirely based on a double-track. In this case all single-track sections existing on the line (between Dazio and Arco Felice stations), will be therefore doubled and no modification will be brought to the current signalling system.
- 2) Intervening both on the infrastructure layout and the signalling system. In this case instead, after having entirely doubled the line (as described in the previous scenario), the signalling system will be upgraded to an ETCS level 1 but not modifying the current block section layout (i.e. their lengths and positions).

A classical “what-if” approach has been employed for evaluating the effects induced in terms of capacity by these two intervention solutions. Actually, also the non-intervention scenario (i.e. the current configuration of the system) has been simulated to firstly determine which is the current state of the network, and then to estimate the increase in terms of capacity that each one of the considered intervention scenarios can

supply, with respect to the current layout. Currently, the service is operated with a headway of 10 minutes and therefore six train runs per direction, as illustrated in the simulated timetable shown in Figure 91.

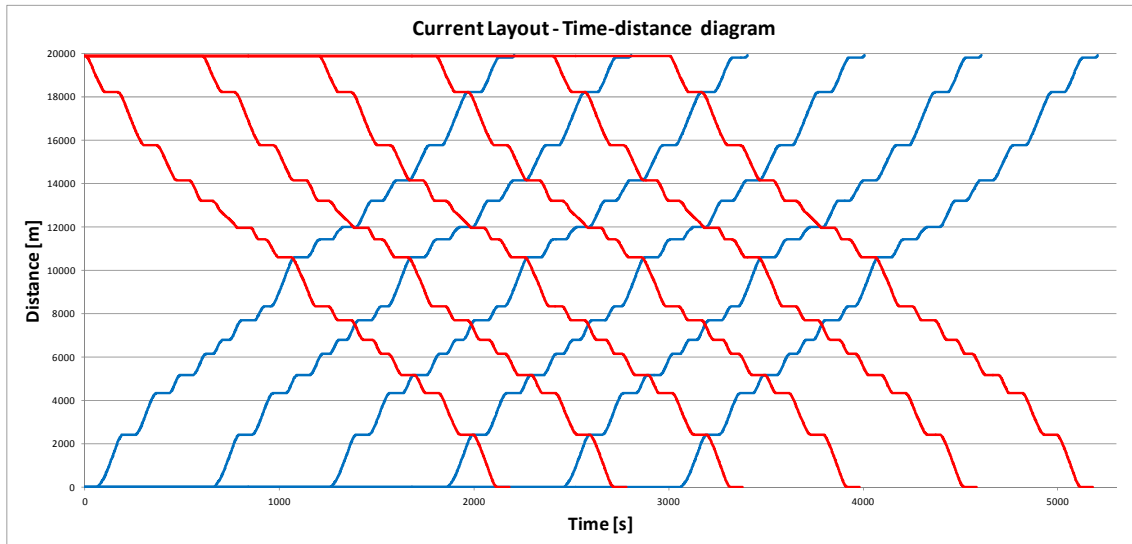


Figure 91. Time-distance train trajectories within the non-intervention scenario (current network layout).

The aforementioned module addressed to calculate the signal headway, has estimated for the current layout, a minimum line headway of about 8 minutes and 40 seconds.

Then the intervention scenario 1 has been implemented in the microscopic model developed, and the corresponding signal headway has been determined by the apposite module. Specifically within this scenario the service headway must be higher than 7 minutes and 40 seconds, since the signal headway estimated for the Montesanto-Torregaveta direction (“Even” track) is 7 minutes and 4 seconds (424 s) while for the opposite direction (“Odd” track) it is just equal to 7 minutes and 40 seconds (460 s). Therefore it is clear that this solution cannot be adopted to meet the required service headway of 5 minutes.

Successively the second intervention scenario has been simulated, and the module dedicated to the calculation of the minimum line headway, has estimated for the “even” track (i.e. Montesanto-Torregaveta direction) a signal headway of 4 minutes and 44 seconds (284 s), while for the “odd” track (i.e. the opposite direction) a signal headway of 4 minutes and 51 seconds (291 s). Therefore this solution is compatible with the required capacity level. Figure 92 illustrates time-distance train trajectories within the intervention scenario 2 assuming 5 minutes as service headway.

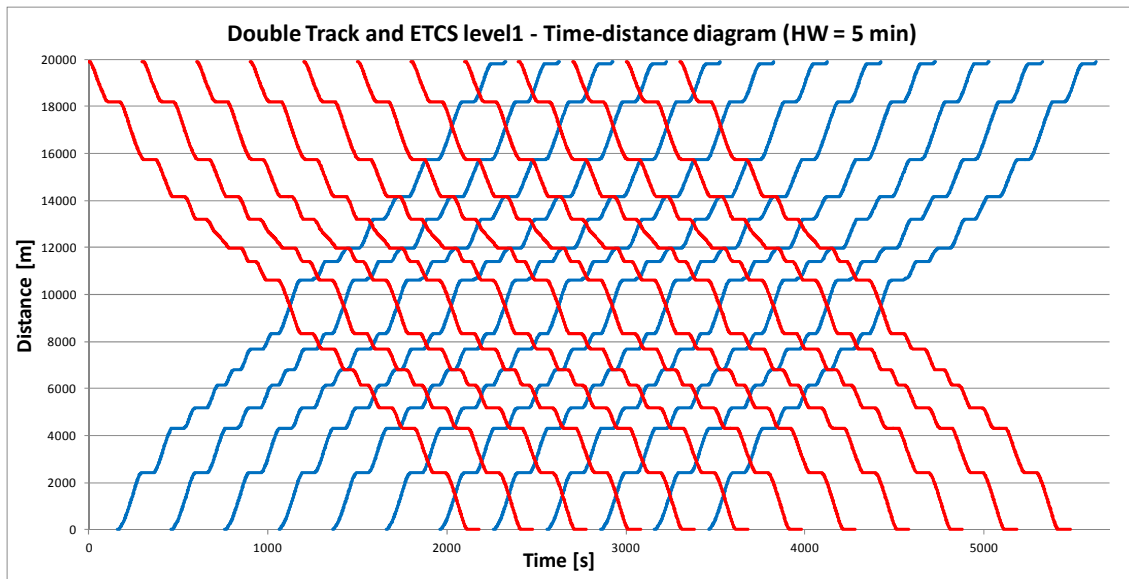


Figure 92. Time-distance train trajectories within intervention scenario 2, operated according to the objective service headway of 5 minutes.

Table 5 summarizes the described results showing for each intervention scenario the minimum service headway that can be delivered and the corresponding increase of network capacity with respect to non-intervention scenario.

Scenario	Signal headway[s]	Capacity increase [%]
Non-Intervention	520	0
Scenario 1	469	10
Scenario 2	291	44

Table 5. Values of the minimum line headway for each intervention scenario and corresponding increases in terms of capacity with respect to non-intervention scenario.

5.3. Designing an equi-block signalling layout maximizing economic efficiency of investment costs for a MRT line.

As shown in the previous application, the minimum service headway that a railway operator can deliver on a certain network strongly depends on the technology relative to the signalling system implemented as well as on its layout (e.g. block section lengths and positions). Therefore if an increase of passenger demand is foreseen and a consequent increase of network capacity is necessary, infrastructure manager will need to carry out “re-signalling” works addressed to upgrade and improve the current signalling system, if this one is incompatible with the required service headway. However, as also described within Chapter 3, the layout of signalling system is usually designed following an approach which tends to maximize the technological efficiency

of the system. In particular the maximum technological efficiency is reached when the length of block sections is equal to the minimum value that it can assume to assure safety conditions during movement, L_{min} . As already known, this minimum value L_{min} is just coincident with the maximum braking distance that a rail vehicle can perform on that track, assuming the lowest possible value of the deceleration rate (i.e. the service braking rate within the “worst-case” scenario). In fact in literature (Gill and Goodman 1992, Chang and Du 1998, 1999, Ke et al. 2009), the maximum braking distance is calculated for a certain track and a certain signalling system type. Then the length of block sections (considering an equi-block layout) is set to L_{min} and for this configuration a single deterministic simulation run is launched to calculate the corresponding signal headway (using the “blocking time” theory) and verify if this configuration satisfies the required value of service headway. Actually, for safety reasons the length of block sections is usually set to $L_{min}+s$, where s is a constant which constitutes a safety margin. Anyway this design approach, guarantees that the capacity level obtained implementing a certain signalling system type, is just the maximum capacity level which that kind of signalling technology could ever provide for the considered network. For this reason such approach can be called as “maximum technological efficiency”.

However, here another design criterion is proposed which tends to maximize instead the economic efficiency of the investment and not just the technological efficiency of the signalling system. Therefore, let h_{obj} be the value of the “objective” service headway that the infrastructure manager has to guarantee to satisfy the foreseen level of demand, the mentioned criterion aims at identifying a signalling configuration which assures the attainment of such headway value (i.e. the signal headway h corresponding to this configuration must be $h \leq h_{obj}$) and contemporarily minimizes the investment costs. In the case of signalling system is easy to understand that investment costs C are minimized when the number of block sections N is minimized, since it is possible to define a linear relationship between these two variables, where the angular coefficient is just the unit cost of the block section c_k for a certain type of signalling system k , therefore $C = c_k \cdot N$. If an equi-block signalling layout is considered, investment costs are minimized when the length of block sections is maximized. Therefore in this latter case, the layout of signalling system which maximize the economic efficiency of the investment can be obtained solving a “black-box” optimization problem where the objective function to maximize coincides with the length of equi-block sections L , while

constraints impose that the signal headway h corresponding to that block section length must be compatible with the “objective” headway h_{obj} . This optimization problem is represented at (50):

$$\begin{aligned} & \text{Max } L \\ & \text{s.t.} \\ & \begin{cases} L_{\min} \leq L \leq L_{\max} \\ h = \mathbf{M}(L) \\ h < h_{obj} \end{cases} \end{aligned} \quad (50)$$

where L_{\min} is the maximum braking distance of the rail vehicle on the considered network (it also depends on the kind of signalling system implemented), L_{\max} represents the maximum length allowed by the track (in particular depends on the minimum distance between two consecutive stations), $\mathbf{M}(\cdot)$ is the microscopic model.

As said before (and clearly highlighted in (50)), this is a “black-box” optimization problem since to obtain the value of the signal headway h corresponding to a certain value of L belonging to the feasible domain, it is firstly necessary to launch a simulation run of the microscopic model and then calculate the minimum signal headway through the “blocking time” theory (illustrated in Chapter 2). To this purpose the microscopic model has been integrated within an optimization loop as illustrated in Figure 93. Practically the microscopic model developed has been interfaced with the C++ API module of the optimization software “LINDO” (LINDO Systems Inc.) through which is possible to solve non-linear “black-box” optimization problems. In particular the optimization algorithm used to solve such problem is the so called OptQuestMultiStart (Ugray et al. 2002) since previous experiences (Ciuffo, Punzo, Quaglietta 2011) showed its efficiency in finding the global optimum independently from the type of objective function employed. Such optimization framework has been applied to the real case of the Cumana line of Naples (Figure 86), supposing that a re-signalling intervention was needed, after having doubled the entire line. To this aim, two optimization problems have been solved to find the optimal signalling layout: one for the “even” track (Montesanto-Torregaveta direction) and one for the opposite track (Torregaveta-Montesanto direction).

Moreover this study has investigated two different kind of signalling systems:

- Three-aspect signalling system based on coded track circuit,

- ETCS level 1.

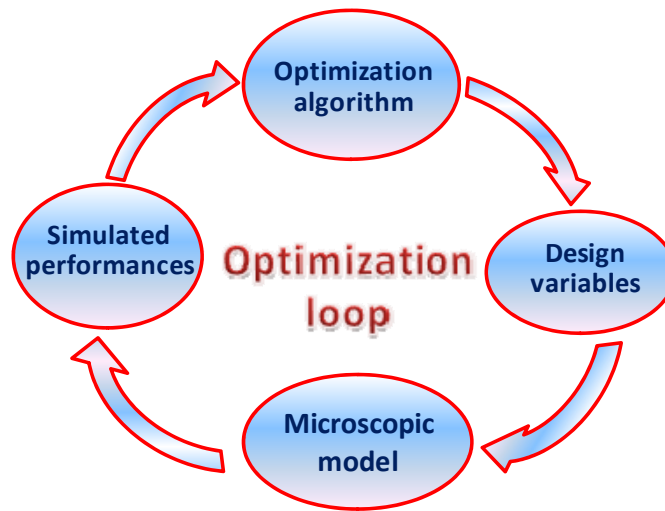


Figure 93. Scheme of the “black-box” optimization loop realized

This means that for each one of the two aforementioned signalling technologies the corresponding optimal layout has been identified. Then once that the investment costs relative to such configurations have been estimated, an economic comparison with the layout which maximizes the technological efficiency (i.e. using a block section length of $L_{min}+s$), has been performed.

Figure 94 and Figure 95 show for the “even” track of the Cumana line the plot of the relationship $h = \mathbf{M}(L)$ between the equi-block section length L and the corresponding signal headway h , as returned by the developed microscopic model $\mathbf{M}(\cdot)$ (and in particular by the module for calculating the signal headway through applying the “blocking-time” theory). In particular Figure 94 shows the plot of such relationship when a three-aspect system based on coded track circuit is implemented while Figure 95 illustrates the same relation when an ETCS level 1 is installed. As can be clearly seen, such functions have some discontinuities where the value of h , first increases linearly with L and then jumps to higher values where a different linear relationship is observed again. These jumps mean that the critical section (i.e. the section that dictates the minimum signal headway and therefore where the blocking times of two consecutive trains overlap) has changed and therefore a different section of the network influences the minimum line headway.

Considering the plot of the function $h = \mathbf{M}(L)$, between the signal headway h and the equi-block section length L for a certain signalling technology, it is possible to represent

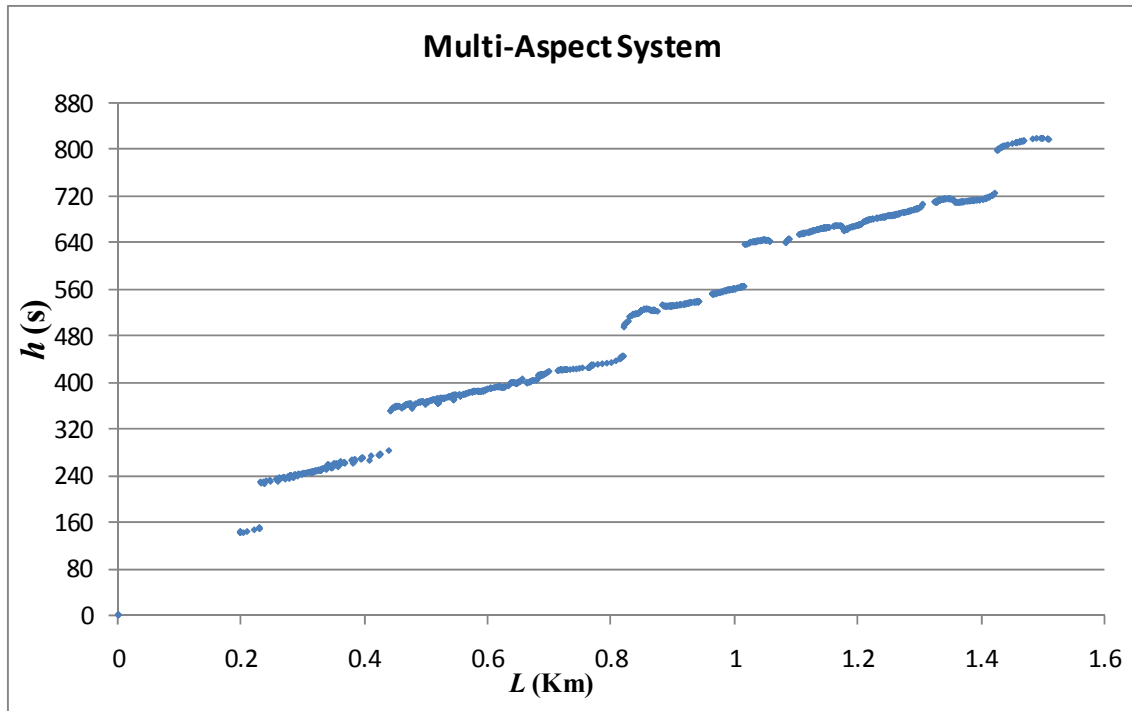


Figure 94. Plot of the relation between the equi-block length L and the corresponding signal headway h for a three-aspect signalling system on the Cumana line ("Even" track: Montesanto-Torregaveta direction).

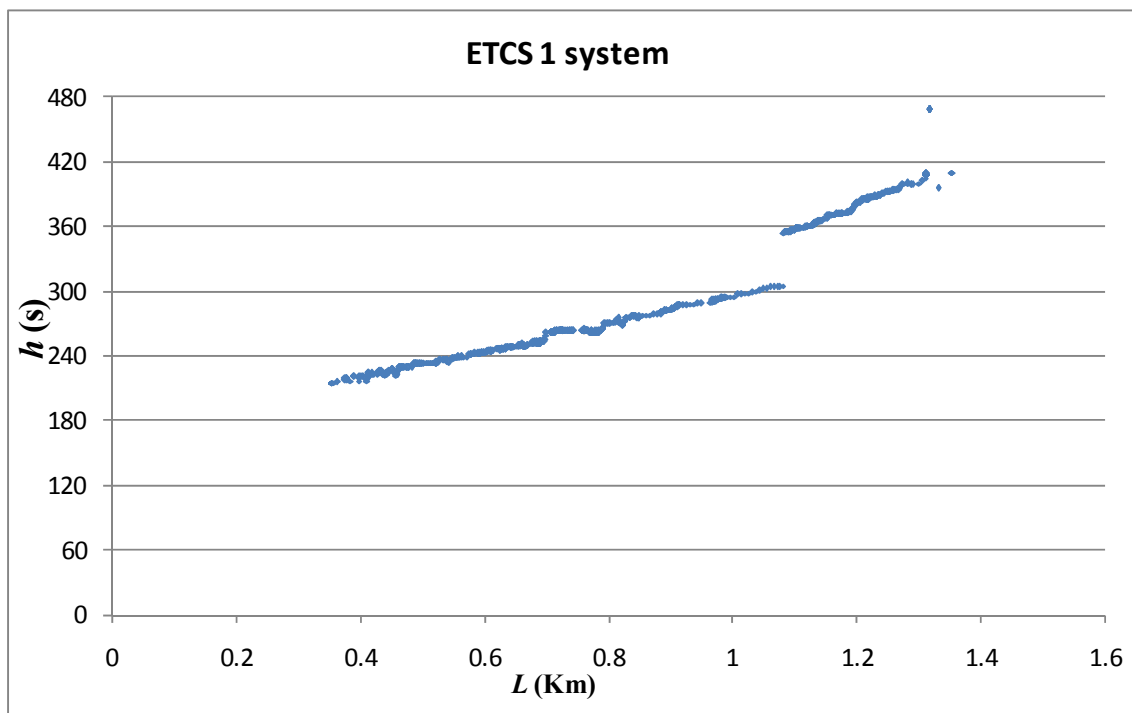


Figure 95. Plot of the relation between the equi-block length L and the corresponding signal headway h for an ETCS level 1 signalling system on the Cumana line ("Even" track: Montesanto-Torregaveta direction).

graphically both the design solution which maximizes the technological efficiency and the solution which instead maximizes the economic efficiency of the investment, i.e. the

solution of the “black-box” optimization problem reported at (50). For example if a multi-aspect signalling system is considered, the plot of the function $h = \mathbf{M}(L)$ has the shape of the diagram showed in Figure 94. Supposing that the infrastructure manager needs to guarantee a service headway $h_{obj} = 320$ s (to satisfy a certain level of passenger demand), the comparison between the design solutions returned by the two aforementioned approaches is illustrated in Figure 96.

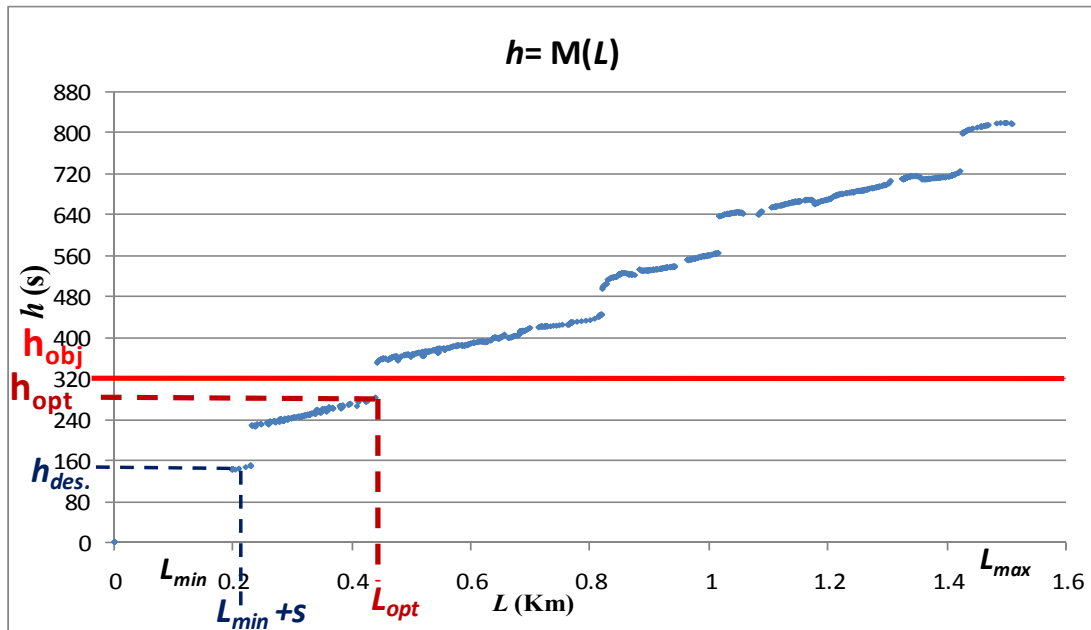


Figure 96. Comparison between the design solution which maximizes the technological efficiency ($h_{des.} - L_{min} + s$) and the design solution which maximizes the economic efficiency of the investment ($h_{opt.} - L_{opt.}$).

As can be seen the design length of the equi-block layout which maximizes the technological efficiency is $L_{des.} = L_{min} + s$, which guarantees a minimum line headway of $h_{des.}$, that in turn practically coincides with the minimum value of signal headway that a certain kind of signalling technology can ever guarantee on the considered network. The optimal solution of the “black-box” optimization problem $L_{opt.}$, and the corresponding signal headway $h_{opt.}$ represent instead the characteristics of the signalling layout which maximizes the economic efficiency of the investment. In fact, in this case the residual capacity (which can be represented by the difference $h_{obj} - h_{opt.}$) is certainly smaller than the residual capacity offered by the previous design approach ($h_{obj} - h_{des.}$), but this configuration allows to obtain a larger length of the equi-block section layout, inducing a strong reduction of investment costs, since a lower number of block sections will be necessary.

As said before a practical application of the proposed design framework has been carried out for the “Cumana” line, supposing that a re-signalling intervention was needed after having doubled the entire line, and the infrastructure manager needs to guarantee a minimum line headway $h_{obj} = 240$ s. Moreover the lower bound L_{min} and the upper bound L_{max} of the equi-block section length have been respectively set to:

- $L_{min} = 0.198$ Km, $L_{max} = 1.5$ Km for the three-aspect signalling system and
- $L_{min} = 0.35417$ Km, $L_{max} = 1.5$ Km for the ETCS level 1 signalling system

where L_{min} represent the maximum braking distance of a train on the track for the considered type of signalling system (calculated according the method illustrated in Gill, Goodman 1992), while L_{max} depends on the physical characteristics of the network (and in particular on the minimum distance between two consecutive stations). Solutions of the “black-box” optimization model reported at (50) have been obtained respectively for the “three-aspect” and the ETCS level 1 systems for both the “Even” (Montesanto-Torregaveta direction) and the “Odd” (Torregaveta-Montesanto direction) track of the line. These results are reported in Table 6. In particular, the investment costs relative to each one of the two signalling technologies have been estimated considering average costs deriving from previous market analyses regarding these kinds of systems.

	Three aspect system		ETCS level 1 system	
	Even Track	Odd Track	Even Track	Odd Track
L_{opt} [Km]	0.292122	0.345271	0.599012	0.697844
N° block sections	68	58	33	29
Cost [€]	9,962,326	8,497,278	9,269,806	8,146,194
Total Cost [€]	18,459,605		17,416,000	

Table 6. Lengths of the equi-block section layout which maximizes the economic efficiency of the investment for both the Three-aspect and the ETCS level 1 systems on the Cumana line (with $h_{obj} = 240$ s)

Economic advantages returned by this design approach can be quantified comparing the relative investment costs with those corresponding to the layout which instead maximizes the technological efficiency of the system (reported in Table 7). This comparison is summarized in Table 8 where it is highlighted that for the considered case-study the adoption of the proposed design approach leads to a reduction of investment costs of 34% for the three-aspect signalling system and 42% for the ETCS level 1. In addition, it is immediate to see that adopting this design approach the investment costs relative to the three-aspect system is higher than the one required to

install the ETCS level 1, although this latter needs the installation of additional components such as transponders and train on-board systems (e.g. on-board computers).

	Three aspect system		ETCS level 1 system	
	Even Track	Odd Track	Even Track	Odd Track
L_{opt} [Km]	0.2100	0.2100	0.3750	0.3750
N° block sections	95	95	53	53
Cost [€]	13,917,956	13,917,956	14,887,871	14,887,871
Total Cost [€]	27,835,912		29,775,742	

Table 7. Lengths of the equi-block section layout which maximizes the technological efficiency of the investment for both the Three-aspect and the ETCS level 1 systems on the Cumana line (with $h_{obj} = 240$ s)

This is clearly due to the fact that since the ETCS level 1 system controls the braking curve of trains (see Chapter 2), it does not need to have an empty block section between two consecutive trains, as instead the three-aspect system does. For this reason the ETCS level 1 can guarantee the “objective” service headway $h_{obj} = 240$ s (considered in the case-study) with a block section length L_{opt} that is longer than the one required instead in the case of the three-aspect system, leading to a strong reduction in the number of block sections and therefore to lower investment costs.

System Type	Cost ($L_{min} + s$) [€]	Cost (L_{opt}) [€]	%Saving
Three-aspect system	27,835,912	18,459,605	34
ETCS Level 1 system	29,775,742	17,416,000	42

Table 8. Comparison between the investment costs induced by the two different design approaches respectively for the three-aspect and the ETCS level 1 systems on the Cumana line (with $h_{obj} = 240$ s)

5.4. Identifying the equi-block signalling layout which guarantees the best “trade-off” between users’ satisfaction and investment costs for a MRT line.

As already said in the previous sections, during the construction of new railway infrastructures or re-signalling works relative to pre-existing systems, designers of signalling systems have the hard task of identifying the signalling technology as well as its layout which is able to guarantee the “objective” service headway h_{obj} required by the infrastructure manager to meet the foreseen level of demand. In turn, since the installation of a new signalling system is a very expensive work, it could be convenient for the infrastructure manager to increase the service headway h_{obj} of a small value ε

which is not perceptible during service by passengers, in order to reduce investment costs when designing the signalling layout according to the “maximum economic efficiency” approach (illustrated in the previous section). In fact using this design criterion it could happen that a small (and undetectable from passengers) increase ε of the objective service headway, can lead to significant reductions in investment costs. In fact as illustrated in Figure 97, if the signalling system is designed considering a service headway equal to $h_{obj} + \varepsilon$ instead of h_{obj} , the length of the equi-block section $L_{opt,\varepsilon}$ is larger than the one corresponding to the objective headway L_{opt} , inducing therefore a reduction of investment costs. Anyway, the corresponding signal headway $h_{opt,\varepsilon}$ is higher than the “objective” service headway h_{obj} , and the larger this difference is the lower passenger’s satisfaction will be, since the higher passenger’s generalized cost will be. Here in fact user’s generalized costs are used as a measure of their satisfaction, in accordance with classical literature on transportation demand (Cascetta 2009).

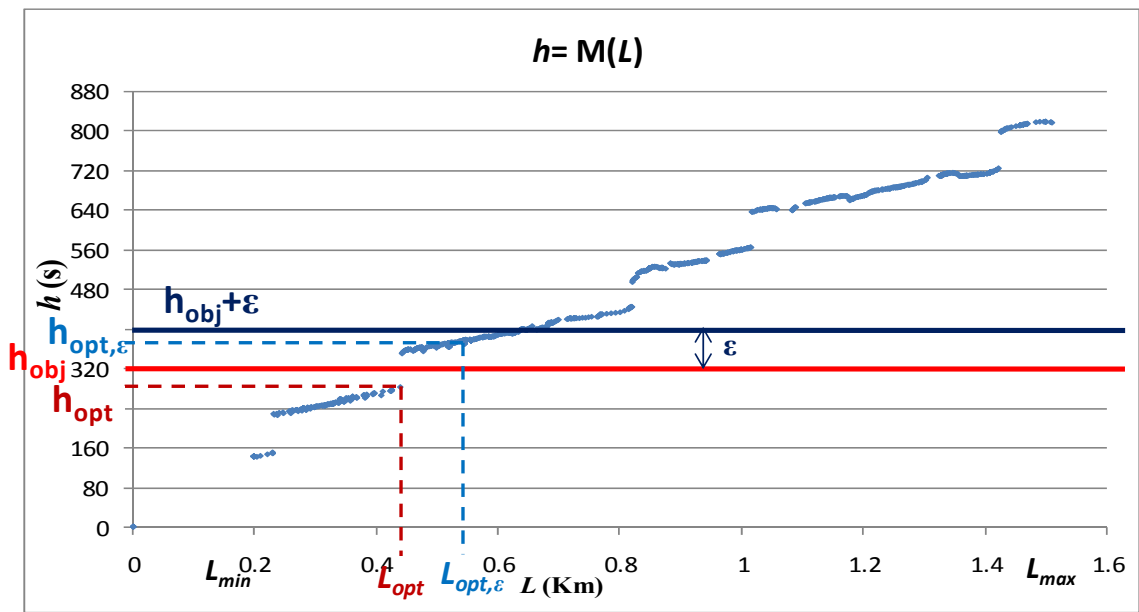


Figure 97. Signalling layout which maximizes the economic efficiency of the investment: comparison between the layout obtained designing with respect to h_{obj} and the one obtained instead considering $h_{obj} + \varepsilon$.

However, when increasing the “objective” service headway of a small value ε , from a side the investment costs for installing the signalling system are decreased, but from the other side the user’s generalized costs are increased, because for example the average waiting times at stations will be higher. Hence, the infrastructure manager could be interested in finding the signalling layout which offers the best “trade-off” between the reduction of investment costs and the increase of passenger’s generalized costs (i.e. the

reduction of user's satisfaction). In other words, a value of ε should be identified, to which corresponds an (optimal) equi-block length $L_{opt,\varepsilon}$ and a related signal headway $h_{opt,\varepsilon}$, that minimize the value of a weighted cost $C_{w,\varepsilon}$ which considers both investment costs of the signalling system and the passenger's generalized costs. In particular for a certain value of ε , such weighted cost can be expressed as illustrated at (51):

$$C_{w,\varepsilon} = Cost(L_{opt,\varepsilon}) \cdot \varphi(User's Gen. Cost_{\varepsilon}(h_{opt,\varepsilon}, h_{obj})) \quad (51)$$

where $Cost(L_{opt,\varepsilon})$ is the investment cost to install a signalling system with an equi-block section length of $L_{opt,\varepsilon}$, while the weight $\varphi(User's Gen. Cost_{\varepsilon}(h_{opt,\varepsilon}, h_{obj}))$ is a function of the passengers' generalized cost (therefore of their satisfaction) that in turn depends on the difference between the signal headway $h_{opt,\varepsilon}$ and the objective service headway h_{obj} . However given a certain value of ε , it is possible to identify $L_{opt,\varepsilon}$ and the corresponding signal headway $h_{opt,\varepsilon}$, solving the “black-box” optimization problem presented at (50) and imposing $h_{obj} + \varepsilon$, as upper bound of the last constraint, instead of h_{obj} , as reported at (52):

$$\begin{aligned} &Max \ L \\ &s.a. \\ &\begin{cases} L_{min} \leq L \leq L_{max} \\ h = \mathbf{M}(L) \\ h < h_{obj} + \varepsilon \end{cases} \end{aligned} \quad (52)$$

In particular such problem must be solved for different values of ε belonging to the range $[\varepsilon_{min}, \varepsilon_{max}]$. Once that for each one of the considered values of ε , both $L_{opt,\varepsilon}$ and the corresponding signal headway $h_{opt,\varepsilon}$ have been found, it is possible to calculate the relative weighted cost $C_{w,\varepsilon}$ and therefore identify the value of ε for which such cost is minimized.

The design procedure described above has been applied to the real case-study of the “Cumana” line, supposing as in the previous section, that the infrastructure manager has to guarantee a service headway $h_{obj} = 240$ s. Thirteen different values of ε have been investigated considering $\varepsilon_{min} = 0$ s and $\varepsilon_{max} = 120$ s, with a step of 10s. Therefore the considered values of ε are: 0s, 10s, 20s, 30s, 40s, 50s, 60s, 70s, 80s, 90s, 100s, 110s, 120s. Then for each one of these values the “black-box” optimization problem (52) has been solved and the optimal solution $L_{opt,\varepsilon}$ as well as the corresponding signal headway $h_{opt,\varepsilon}$, have been found. After that, the relative value of the weighted cost $C_{w,\varepsilon}$ has been

calculated. Specifically, for the sake of simplicity the weight $\varphi(\text{User's Gen.Cost}_\varepsilon(h_{opt,\varepsilon}, h_{obj}))$ of the weighted cost, has been simply represented as the ratio between the signal headway $h_{opt,\varepsilon}$ and the objective service headway h_{obj} :

$$C_{w,\varepsilon} = \text{Cost}(L_{opt,\varepsilon}) \cdot (h_{opt,\varepsilon}/h_{obj}) \quad (53)$$

Hence, only a qualitative trend of $C_{w,\varepsilon}$ with respect to ε has been given. Anyway, this has been enough to identify the value of ε for which the best “trade-off” between investment costs and users’ cost is reached.

Such trade-off has been found both for a three-aspect signalling system and for an ETCS level 1 system. Moreover as in the previous section, the optimal layouts have been identified for both the “even” (Montesanto-Torregaveta direction) and the “odd” (Torregaveta-Montesanto direction) track, supposing that a preliminary work addressed to double the entire line was carried out. However results relative to the entire line are reported in Table 9 a and b, where the values of the weighted cost $C_{w,\varepsilon}$, as described at (53) has been reported for the different values of ε as well as for the two different kinds of signalling systems taken into account.

Three-Aspect TOTALS				ETCS LEVEL 1 TOTALS			
ε [s]	N° block sections	$C(L_{opt,\varepsilon})$ [€]	$C_{w,\varepsilon}$ [€]	ε [s]	N° block sections	$C(L_{opt,\varepsilon})$ [€]	$C_{w,\varepsilon}$ [€]
0	126	18,459,605	18,459,605	0	62	17,416,000	17,416,000
10	113	16,555,042	17,213,093	10	56	15,730,581	16,355,590
20	105	15,383,004	16,512,312	20	54	15,168,774	16,339,204
30	97	14,210,966	15,867,691	30	50	14,045,161	15,407,542
40	92	13,478,442	15,611,307	40	49	13,764,258	15,388,815
50	89	13,038,927	15,536,224	50	46	12,921,548	15,613,538
60	86	12,599,413	15,282,282	60	41	11,517,032	14,396,290
70	86	12,599,413	15,282,282	70	38	10,674,323	13,658,919
80	86	12,599,413	15,282,282	80	38	10,674,323	13,658,919
90	86	12,599,413	15,282,282	90	38	10,674,323	14,150,500
100	86	12,599,413	15,282,282	100	37	10,393,419	14,637,399
110	86	12,599,413	15,282,282	110	36	10,112,516	14,452,471
120	80	11,720,384	17,531,741	120	36	10,112,516	14,915,961

Table 9. Weighted costs $C_{w,\varepsilon}$ for different values of ε , relative to the case of the three-aspect signalling system (a) and the ETCS level 1 (b).

Such results have been reported in a graphical form respectively in Figure 98 for the three-aspect system and Figure 99 for the ETCS level 1. As can be seen, in the case of the three-aspect system the weighted cost is minimized when $\varepsilon = 60$ s. In fact designing the signalling layout (according a “maximum economic efficiency” approach) with respect to $h_{obj} + \varepsilon = 240+60=300$ s and not just with respect to $h_{obj}=240$ s, it is possible

to save about 6 millions of € in investment costs, since in this case only 12,599,413 € will be necessary (see Table 9 a) instead of 18,459,605 € (needed for $\varepsilon=0$ and therefore for h_{obj}). Anyway, if a deviation of 60s from the objective headway h_{obj} , seems to be too large (because for example it is quite perceptible from passengers), it is always possible to define a smaller value of ε and identify the local minimum of the weighted cost function within the region $[0 \varepsilon]$. For example if a maximum value of 30s is accepted for ε , it is immediate to see from Figure 98, that the local minimum of the $C_{w,\varepsilon}$ function within the range $[0s \ 30s]$ is obtained just in correspondence of 30s. Moreover considering this value of ε , a consistent reduction of investment costs is also obtained since in this case 14,210,966 millions of € are needed instead of 18,459,605 €, leading to a saving of more than 4 millions of €.

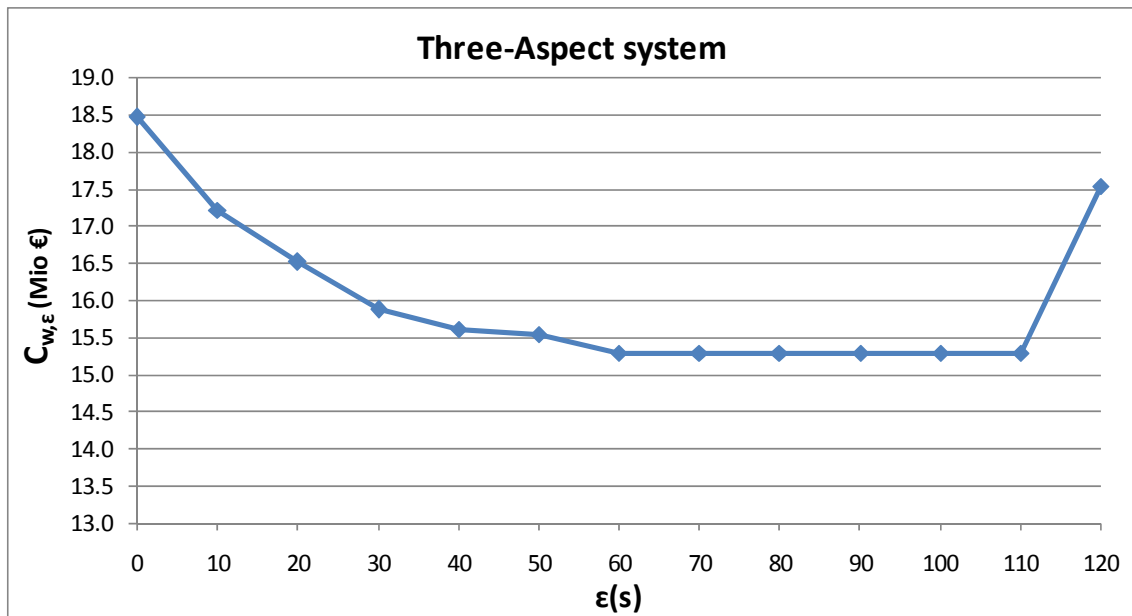


Figure 98. Weighted cost $C_{w,\varepsilon}$ for different value of ε , for the three-aspect signalling system.

The same consideration can be made for the ETCS level 1. In particular in this case the minimization of the weighted cost function $C_{w,\varepsilon}$ is obtained for $\varepsilon = 70s$, to which corresponds a total saving in investment costs of about 7 millions of €, since it will be required only 10,674,323 € instead of 17,416,000 € (see Table 9 b). However if a deviation of 70s from the objective service headway is considered to be too large, it is possible to determine a smaller value of ε and identify the local minimum of the weighted cost function within the region $[0 \varepsilon]$. For instance, if such a value has been fixed to 50s, it is immediate to see from Figure 99, that within the range $[0s \ 50s]$ the $C_{w,\varepsilon}$ function has a local minimum for $\varepsilon = 30s$, to which a total saving of more than 3

millions of € is reached, since only 14,045,161 € will be necessary instead of 17,416,000 €.

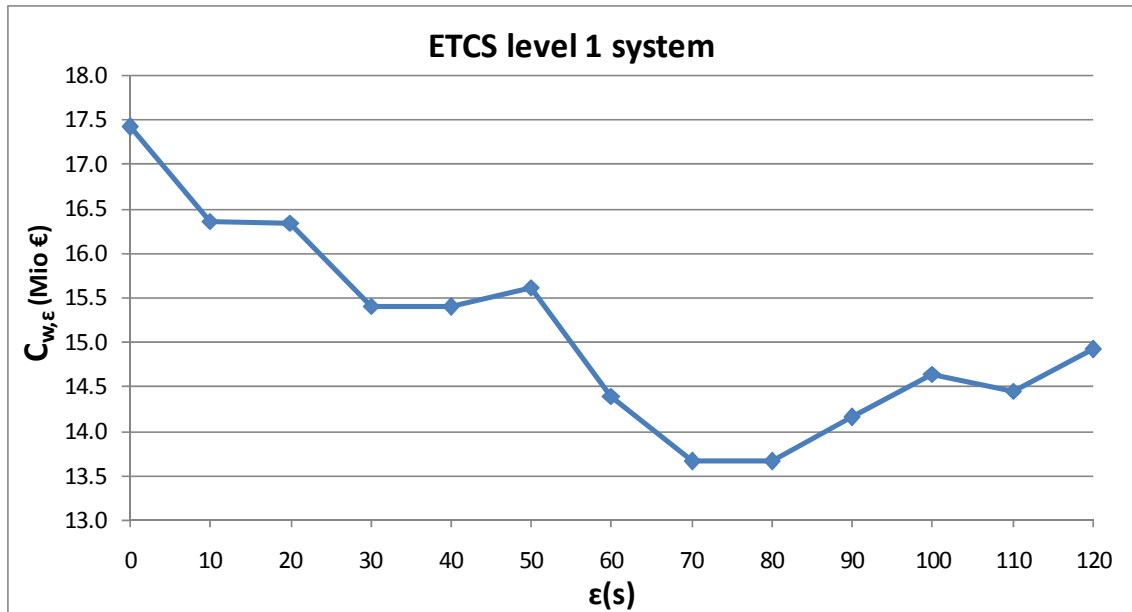


Figure 99. Weighted cost $C_{w,\epsilon}$ for different value of ϵ , for the ETCS level 1 signalling system.

5.5. A model for supporting RAM analysis: towards a hybrid integration with an “event-driven” mesoscopic model.

Recently the introduction of the norm CENELEC EN 50126, has established the need for European railway actors at each level, to adopt effective infrastructural solutions and/or management procedures aiming at minimizing the effects of components failures on network performances. To this purpose the effectiveness of a certain infrastructural and/or operational intervention must be evaluated since the earliest stages of design, in order to verify if the required levels of service are satisfied during the whole system lifecycle. This problem is in particular faced by suppliers of railway components (e.g. signalling systems, vehicles, interlocking and power supply systems), which have to respect certain service requirements as specified by customers within contract specifications. If such requirements will be disappointed, suppliers will have to pay very heavy economic penalties which are often in the range of several millions of €. For this reason these actors strongly need to verify since designing phases, if the characteristics of their components, and in particular *reliability* (the inverse of the failure rate) and *maintainability* (i.e. the probability of the item to be maintained in a time period under given conditions), as well as recovery strategies identified, are able to satisfy the levels of service *availability* (e.g. punctuality, regularity) requested by customers. To this aim

a specific probabilistic analysis is required, just called R.A.M. (*Reliability, Availability and Maintainability*) analysis where millions Monte-Carlo simulations of railway service must be performed in order to draw a significant number of failure events and estimate their effects on system performances (and in particular on service availability). However to effectively carry out this kind of analysis, it is necessary to employ a simulation model which is contemporarily efficient for performing this large number of simulations, and accurate to precisely evaluate the effects of failure events on the network. Although the microscopic model developed is very accurate and efficient when it is launched on a multi-core computer, several millions of Monte-Carlo simulations are however too much to be performed in an acceptable period of time. Therefore it is necessary to overcome this applicability limit relying on a different kind of model. In particular an “own-built” mesoscopic model has been developed (in the context of another PhD research activity) which considers as input data the failure rates of each component and whose computational efficiency is able to perform millions Monte-Carlo simulations in few minutes, since it is based on a Stochastic Activity Network (SAN) formalism (an evolution of the timed Petri Nets). Anyway, since this model is a “fixed-speed” model and is not able to describe the dynamic evolution of rail vehicles on the track, it lacks in accuracy, especially when transient train dynamics strongly influence system performances (e.g. during congested conditions). Hence, in order to overcome the applicability limits of the two models, a microscopic-mesoscopic model that dynamically integrates the two approaches is proposed, to effectively support the aforementioned type of probabilistic analysis. In particular the integration has not been realized yet, but after a practical evaluation of the trade-off between efficiency of the meso and accuracy of the micro model, an integration strategy has been identified. This strategy will be explained after a brief description of the mesoscopic model and the illustration of results relative to the quantification of differences between the two models in terms of both accuracy and computational efficiency.

5.5.1. Mesoscopic Model

As already said the mesoscopic model has been implemented using the Stochastic Activity Networks (SAN) formalism to represent high-level system behaviours in order to evaluate global impacts of failures on network operation, assessing reliability, availability, maintainability and performability levels of different track layouts and fall-

back strategies (i.e. the set of actions to effect when a failure occurs to restore normal operation or activate degraded operation modes). This is an event-driven multi-train simulation model. Input data can be classified with respect to the modules by which the architecture is composed of:

- *Behavioural module.* The inputs required by this module are: train operation attributes like free-flow train travel times for each block section and dwell times at stations as well as train headway. In addition this module needs information on how signalling system regulates train movements and about fall-back strategies (actions to restore ordinary service after a failure).
- *Component status module.* Input data here required pertain to failure rates and corresponding MTTR (Mean Time To Restore, i.e. the average time to restore ordinary conditions) for critical components (vehicles, signalling system, track equipment, etc.). This module is responsible for the simulation of components transitions between ordinary and degraded status and vice versa.
- *Evaluator module.* This module evaluates system performance indexes which are of key relevance for the investigation. Estimation of such indexes is realized by averaging results of different simulation replications.

Network is modelled as a graph where nodes are block section joints and stations, while links represent the time to cross block sections themselves. Train movement on the track is modelled as a sequence of activities whose durations tally with free-flow train travel times on corresponding track sections. Figure 100 shows in fact the timed Petri-Nets used to represent train movements. In particular such figure represents the case of a train crossing a block section which includes a station. If the protection signal at entrance (node “Sem_Prev”) is green, the train enters the block section (node “Entering Station”), and then after a certain running time (link “Entering Time”) it reaches the station and stops there (node “Stopped”) for a certain time period (link “Stop Time”). After that, if the exit signal of the station is red, the train must wait (node “Waiting for Sem”) until it will be green and after a certain running time (link “Exiting Time”) it will exit the block section (node “Exiting Station”). Simulation goes ahead with the realization of discrete events (e.g. train arrivals or departures to/from block section joints and stations) whose timestamps are dictated by ending/starting times of relative activities. In other words the duration of each activity is assigned to links of the network

and therefore the transition from an event (e.g. the train enter the block section represented by the node “Entering Station”) to another event (e.g. the train is stopped at station, described by the node “Stopped”) lasts just the time necessary to complete the related activity (e.g. the running time needed to pass the block section joint and reach the station, given by the link “Entering Time”).

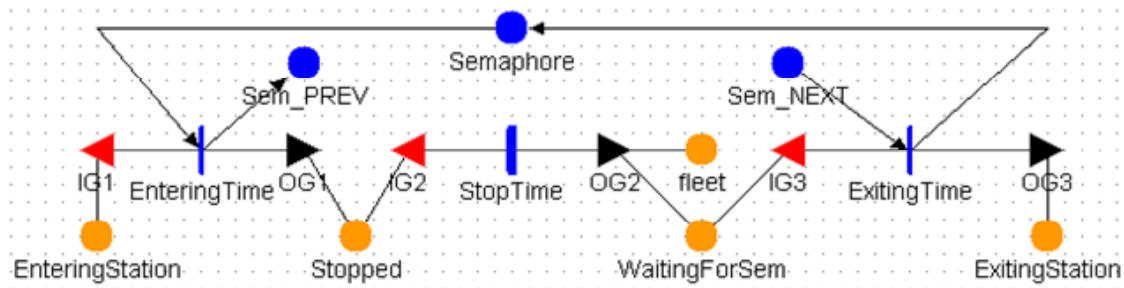


Figure 100. Mesoscopic model based on the SAN formalism: scheme for modelling train movement

Moreover simulation of strategic operations to restore network degraded conditions is possible (e.g. move a broken train to the depot and substitute it with a spare).

Output data provided by this model concern:

- Simulated train arrival/departure times to/from each block section and each station
- Performance parameters (e.g. punctuality, regularity, delays, etc.).

Practical applications revealed the high computing efficiency of such model which is mostly due to the fact that it simulates only main events like train arrivals/departures from sections joints and stations as well as changing of signals aspects, without considering events relative to train acceleration/deceleration phases.

5.5.2. Quantification of differences between the two models in terms of results accuracy and computational efficiency.

In order to identify the integration strategy of the two approaches, the evaluation of the trade-off between results accuracy of the microscopic and computational efficiency of the mesoscopic model has been necessary. To this purpose, both models have been applied to the same case study and respective differences in terms of both accuracy and efficiency have been quantified for different congestion levels on the network. In particular a portion of a simple MRT line has been considered, composed of three stations (A, B and C), and equipped with an ETCS level 1 signalling system having an equi-block section length of 354.17 m. Figure 101 shows the track layout and the

altimetric profile of the considered MRT line portion. Scheduled train headway is 480 seconds (8 minutes) while dwell times at stations are all equal to 30 seconds. The signal headway calculated is 90 seconds. Train attributes used as input data to the microscopic model, are here listed: train length = 80 m, maximum deceleration rate = 1 m/s^2 , jerk rate = 0.75 m/s^3 , total train weight (1 traction unit + 1 wagon) = 103 tons. The actual “tractive effort-speed” curve of the traction unit was considered.

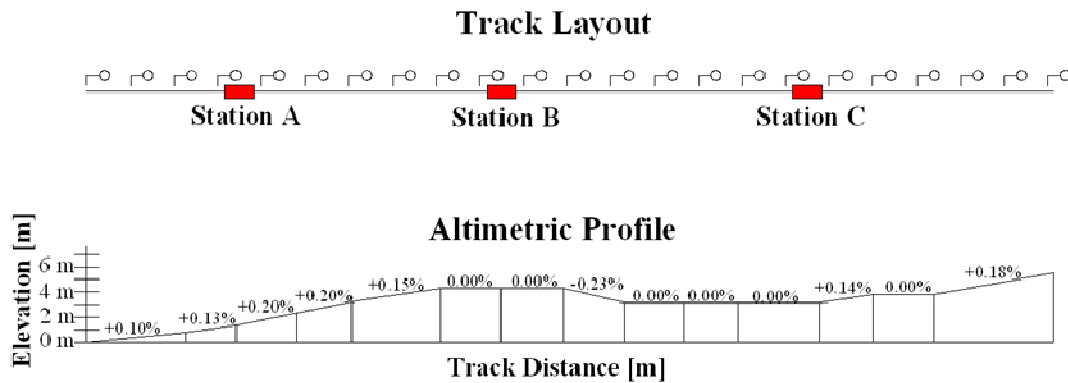


Figure 101. Track layout and altimetric profile of the MRT line portion considered in the case-study.

However to compare mesoscopic and microscopic simulation results, it is necessary that network representations in the two models are consistent. This means that positions of stations and line-side signals as well as lengths of inter-station links and block sections must be the same both in the microscopic and in the mesoscopic model. To this purpose the microscopic network model had been firstly built and then the correspondent mesoscopic representation was consistently extracted from it. Successively to determine mesoscopic input data constituted by free-flow train running times for each block section, a preliminary microscopic simulation of the line with only one train (i.e. without any kind of hindrance to train run) was realized. Once microscopic and mesoscopic models and their respective input dataset had been completely set up, the consistency between the two models was verified through the ascertainment that after simulating an undisturbed 1 hour operation period of the line (using the scheduled train headway), train arrival times at stations, returned as outputs by the two models, were exactly congruent. As mentioned before, differences between microscopic and mesoscopic model outputs arise when unforeseen service disturbances occur in the network, because in this case local and transient train dynamics (e.g. additional

decelerating and accelerating phases due to incoming train conflicts) can be only described by the microscopic one. To highlight and analyse such differences, an original departure delay to the first train at station B was imposed, and correspondent knock-on delays transferred to other trains were measured. In particular a failure to a passenger door closing system was supposed, which forced the first train to be stopped at station B for additional 900 seconds (15 minutes) until the breakdown was repaired. Moreover the possibility for other trains of using alternative paths as well as moving the broken train on a near shunting yard had been excluded, therefore each train had to wait for the re-establishment of ordinary service conditions. The differences between the two models have been evaluated considering different congestion levels and therefore different values of train departure headways, since the less is the departure headway the higher will be the congestion level on the network when a disturbance as the one taken into account is present. Hence, the effects induced on performances by the aforementioned failure at station B have been estimated for 14 different train departure headway values, starting from the theoretical minimum headway of 90s, up to the scheduled headway (480s) with a step of 30s (i.e. 90, 120, 150, 180, 210, 240, 270, 300, 330, 360, 390, 420, 450, 480).

Quantification of differences between the two models in terms of results accuracy

Figure 102a and b, respectively show results given by the microscopic and the mesoscopic model, for each one of the 14 different simulation experiments. In particular for each train run (Train ID) during a simulation period of 1 hour, the corresponding arrival delay at station C has been reported. As expected, knock-on delays decrease when train headway increases. Furthermore when train headway is less than 150 seconds the microscopic model estimates that the considered original delay is not completely recovered within the simulation period of 1 hour, therefore a reduction in line capacity occurs. In particular when train headway is set to its minimum (90 seconds) a line capacity reduction of 26% (from 35 trains/h to 26 trains/h) is assessed. The mesoscopic model instead, in the same case predicts only a capacity reduction of 3% (from 35 trains/h to 34 trains/h) with an overestimation of 31% with respect to capacity value foreseen by the microscopic model. Moreover as Figure 102b shows, the mesoscopic model estimates a capacity reduction, only for the minimum headway (in fact this is the only one curve which does not intersect abscissa axis). Then in order to quantify for each headway value (i.e. for several congestion levels), differences between

outputs returned by the two models, the following percentage deviation indicator ε , was used:

$$\varepsilon = \left| \frac{TotalDelay_{C,meso} - TotalDelay_{C,micro}}{TotalDelay_{C,micro}} \right| \cdot 100 \quad (54)$$

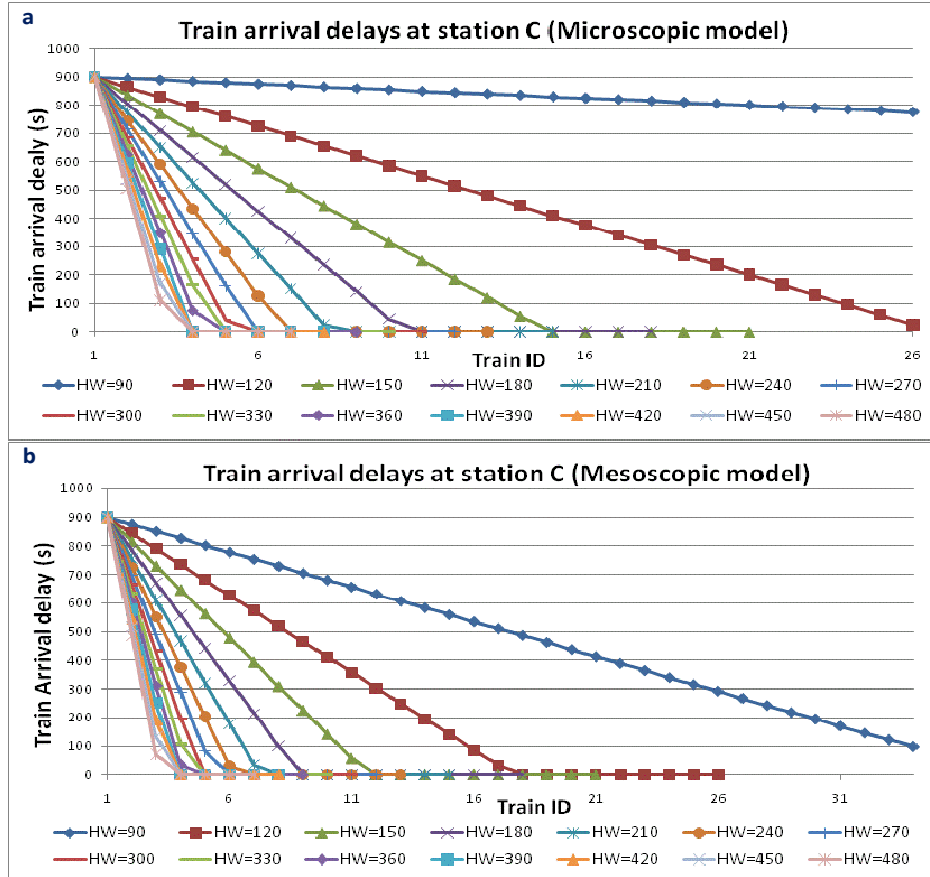


Figure 102. Arrival delays at station C returned by the microscopic model (a) and the mesoscopic model (b), for each train run and for the different train departure headways, within 1 hour simulation period.

This indicator measures the difference between the total train arrival delay at station C (i.e. the sum over all the trains of their respective arrival delays at station C) respectively estimated by the mesoscopic ($TotalDelay_{C,meso}$) and the microscopic model ($TotalDelay_{C,micro}$), normalized with respect to the latter. Figure 103 illustrates how this indicator varies according to different train headway values, and therefore within different conditions of congestion. In particular when train headway is set to its minimum value (high congestion levels), the mesoscopic model underestimates of 40% the total train arrival delay returned by the microscopic one, while for train headways larger than 360 seconds (low congestion levels) such deviation goes down to 5%.

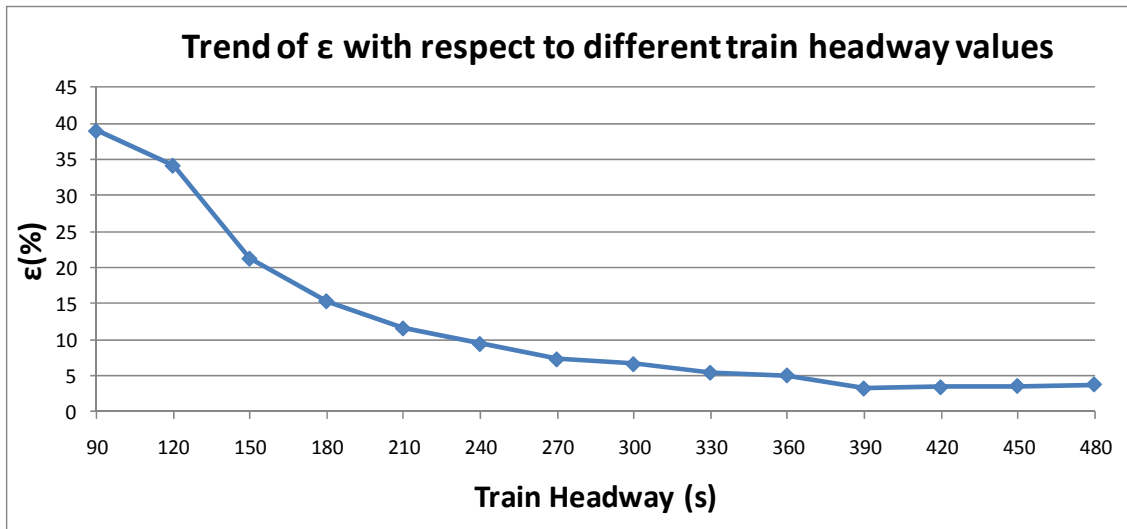


Figure 103. Values of the percentage deviation indicator ε for the different train headway values considered.

Quantification of differences between the two models in terms of computational efficiency

As regards computational efficiency of the two models, in Figure 104 computational times in logarithmic scale are reported for both the microscopic and the mesoscopic model, to realize a single simulation replication (running on a Pentium IV 3.00 GHz processor) for each considered train headway. As can be seen, mesoscopic model processing times are consistently lower than microscopic ones. Furthermore while the microscopic model computing times (using a simulation time-step of 1 second) increase as train headways decrease (because network congestion increases, and the number of trains to simulate in a time-step increases), the mesoscopic model exhibits an opposite trend since the higher is the congestion level on the network the less is the number of events to process in the considered simulation period, because more trains come to a standstill. However the most part of this large difference in terms of computing times is due to the fact that the mesoscopic model simulates only main events such as train departures/arrivals from/at block section joints and stations not considering other events like acceleration or deceleration of trains. Moreover it is a “fixed-speed” model therefore train running times are already known during simulation as they constitute an input of the model itself. In contrast train running times are outputs for the microscopic model which in fact needs to perform for each time step of the simulation period and for each train, a time consuming integration of the Newton’s motion formula, in order to accurately describe train movements on the track. It is clear therefore that such a gap in computing times, will become larger when decreasing the simulation time-step and/or

when increasing the number of trains on the network (i.e. increasing congestion levels), since in this cases the computing times will remain more or less the same for the mesoscopic model, but it will strongly increase for the microscopic model.

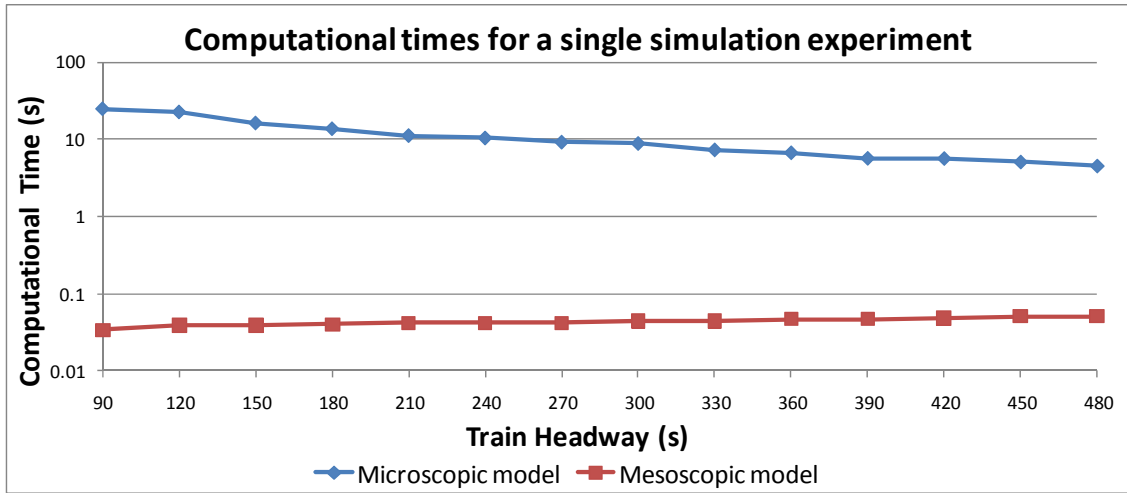


Figure 104. Measured computational times of the microscopic and the mesoscopic model for a single simulation experiment.

5.5.3. Identification of the dynamic integration strategy between the two models.

The outcomes of the comparison between the two models in terms of both results accuracy and computing efficiency, highlight the applicability limits relative to the two approaches and underline the necessity to dynamically integrate the models for effectively supporting RAM analysis. As said before, the quantification of such differences has been necessary to drive the integration strategy and to understand how the two models could be interfaced and above all in which phase each model must intervene, to maximize the efficiency and the accuracy of the integrated approach. In particular such preliminary analysis has led to the integration strategy illustrated in Figure 105. Specifically the microscopic model is firstly launched to perform an initial simulation of the system within undisturbed ordinary conditions, in order to calculate for each train the corresponding free-flow running times. Then such running times are transferred to the mesoscopic model to initialize the lengths of its links. At this point, millions Monte-Carlo simulations of ordinary service are realized by means of the mesoscopic model, to draw stochastic breakdown events according to failure rates specified for each system component as input of the mesoscopic model itself. When a failure event is drawn, its space-time coordinates (which are outputs of the meso model) are transferred to the microscopic model which is in turn activated to perform a

simulation of the correspondent failure scenario and accurately estimate effects of such failure on performances. After that, the mesoscopic model is activated again and the described loop restarts until is reached the total number of simulations needed by the considered probabilistic analysis. Hence, in this case the efficiency of the mesoscopic model is exploited to carry out millions simulations of ordinary service only, since only within nominal and undisturbed conditions the mesoscopic model returns the same outputs of the microscopic model. This latter in turn is activated only when a failure event is drawn, since it is inefficient for realizing a large number of simulations, and its accuracy is exploited only when needed, i.e. when a breakdown is present on the network to precisely describe its repercussions on the system.

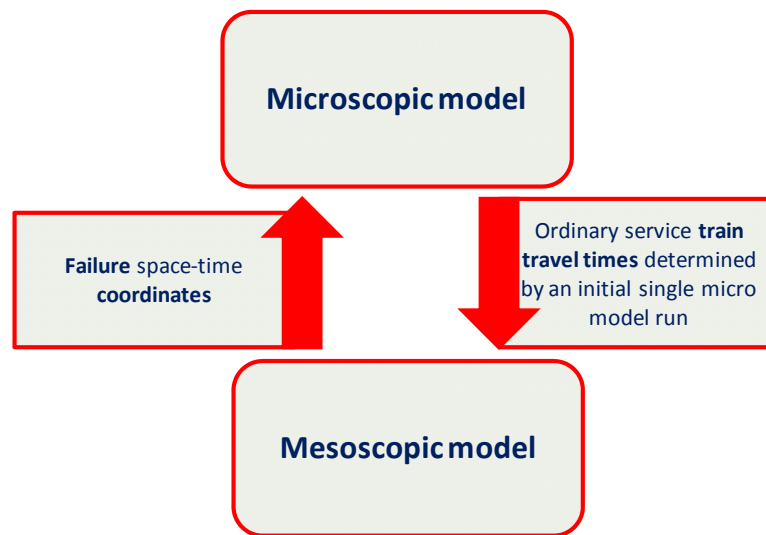


Figure 105. Integration strategy for the dynamic cooperation of the microscopic and the mesoscopic model.

5.6. A model for evaluating effects on both Service Availability and Quality of Service: integration with a module for simulating passenger demand.

Railway actors evaluate the effectiveness of a certain infrastructural and/or operational intervention, mostly considering its repercussions on service availability (e.g. measuring delays, punctuality, regularity, etc.) and in general on network performances, often neglecting its effects on travel demand. However this is mainly due to the fact that railway designers and operators take into account for impacts on customers only implicitly, giving more attentions to the respect of service availability indices as specified within contract requirements. This aspect is in contrast with the concept of railway network, which instead can be considered as a “*demand-oriented*” system, since

apart the improvements of economic and environmental conditions of surrounding areas, it is generally built to content certain travel demand levels. Moreover, in recent years international policies (*European Rail Research Advisory Council* 2007) have been introduced in the field of railway systems, to underline the need of measuring customers' satisfaction levels by means of apposite methodologies and parameters, and above all to address both industry and research activities towards the improvement of railway performances to meet demand growth targets. Therefore the achievement of these aims can be realized only if within decisional processes, railway actors are able to assess impacts of design or strategic solutions not only on service performances but also on quality levels offered to customers. To support this kind of activities an opportune simulation tool is necessary, which not only estimates effects on Service Availability (SA), but explicitly simulates also passenger demand to contemporarily evaluate repercussions on Quality of Service (QoS). To this aim the developed microscopic model has been interfaced with a demand assignment module which considering as input both origin-destination (O/D) matrices of passenger trips and train run performances (given as output by the microscopic model) returns the passenger load for each train run. In turn these outputs are employed to calculate passengers' generalized cost as a measure of QoS delivered to customers.

Specifically the integration carried out between the microscopic model and the demand assignment module is a simple and static input-output link as shown in Figure 106.

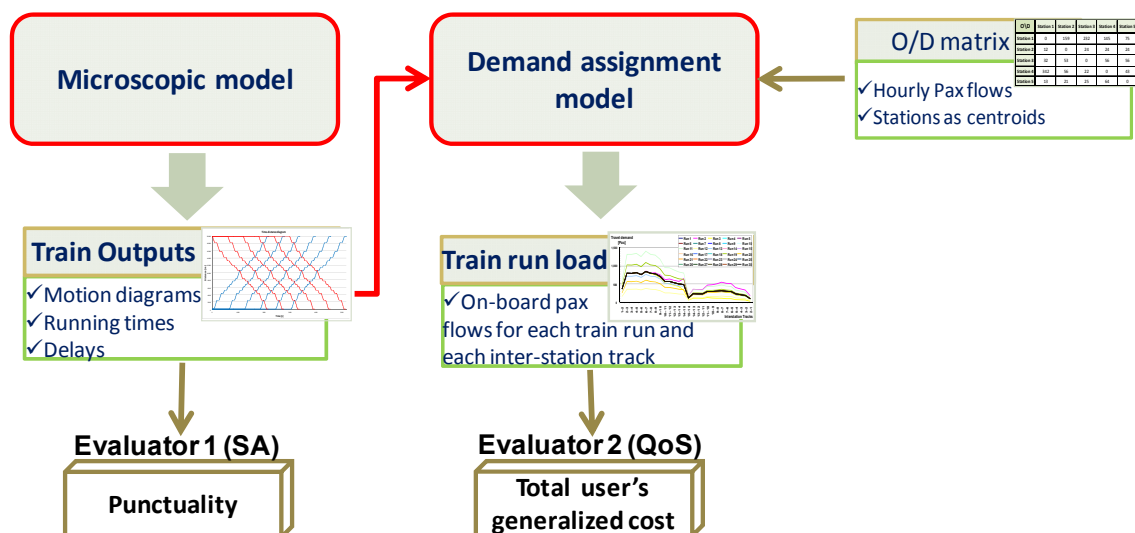


Figure 106. Scheme of the static input-output cooperation between the microscopic model and the demand assignment module.

In particular the microscopic model is firstly launched to predict train performances (e.g. running times, arrival/departure delays, etc.) relative to a certain simulation scenario. Then such performances are usefully employed to calculate SA levels, for example in terms of punctuality index. Successively these train outputs are transferred as inputs to the demand assignment module which considering both boarding and alighting passenger OD matrices (where centroids coincide with stations) referred to a certain time period, determines on-board passengers flows for each train run and each inter-station section, through a deterministic assignment model based on consistency equations (e.g. deterministic network loading). In turn these outcomes are used to calculate the total user's generalized cost as a measure of QoS.

5.6.1. Application to a MRT line

The simulation structure described above has been applied to a real scale case-study of a MRT line, analyzing two different failure scenarios and comparing for each one of these, two recovery strategies in terms of the corresponding effects on both punctuality index (as a measure of SA levels) and total user's generalized cost (as measure of QoS levels). Specifically punctuality index has been here calculated as:

$$Punctuality = (t_s - t_l) / t_s \cdot 100 \quad (55)$$

where t_s is the number of scheduled trips within a certain time period and t_l is the number of lost and delayed trips (i.e. the number of trips which do not arrive or arrive over a delay threshold of 3 minutes at the considered station) calculated over the same time interval.

The user's generalized cost C_i , relative to a single rail passenger for choosing alternative i can be expressed as a linear combination of the K attributes ($X_{K,i}$) concerning that alternative weighted by their respective homogenization coefficients $\beta_{K,i}$, which mostly represent specific costs of the attribute:

$$C_i = \sum_K \beta_{K,i} \cdot X_{K,i} \quad (56)$$

In particular for the case-study presented the average waiting time at station ($X_{wait,i}$) and the on-board running time for reaching station D from station O ($X_{on-board,i,O-D}$) have been considered as attributes of the generalized cost.

$$C_i = \beta_1 \cdot X_{wait,i} + \beta_2 \cdot X_{on-board,i,O-D} \quad (57)$$

Moreover values of the homogenization coefficients (β_1, β_2) relative to both attributes have been set to 5 €/hour.

The considered MRT line has a double-track layout, 15 stations for each direction and it is equipped with an ETCS level 1 signalling system type. Moreover this network involves a depot connected to the track between stations no. 1 and no. 2, a pocket track to store away corrupted trains (between station 7 and 8) as well as two switches to let trains change their path or reverse their direction (Figure 107). Scheduled train headway is set to 6 minutes and train dwell times are all equal to 20 seconds for each station.

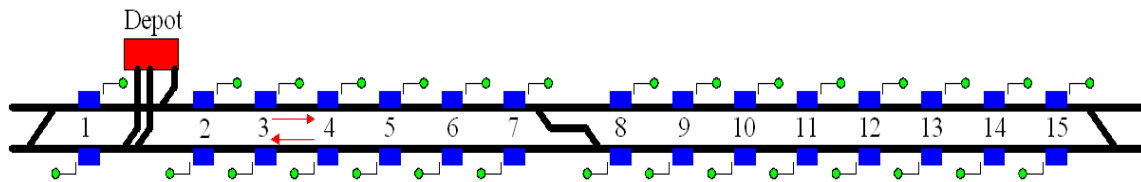


Figure 107. Schematic layout of the considered MRT line.

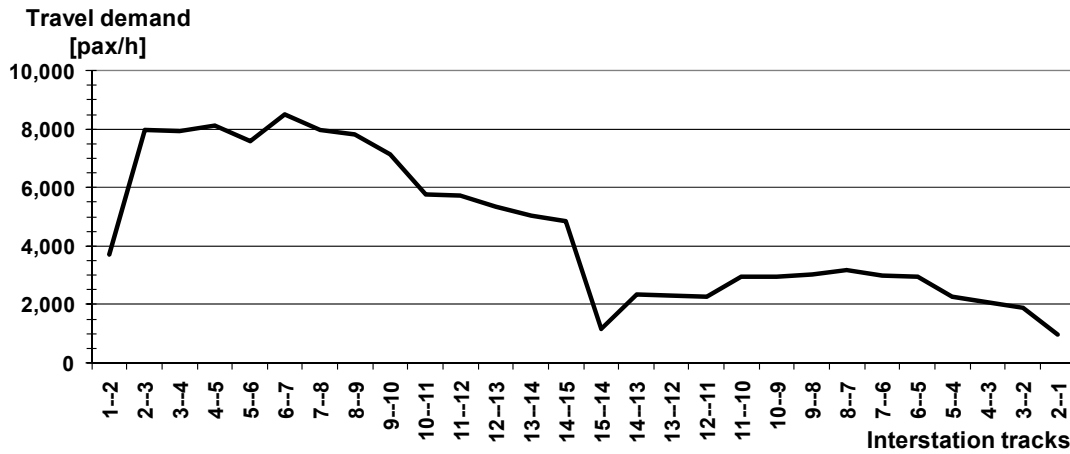


Figure 108. Hourly boarding passengers' flows for each inter-station track.

Figure 108 shows passenger travel demand considered for the line, in terms of on-boarding passengers flows relative to a working day morning peak-hour. Maximum flow for 1-15 direction is 8500 pax/h, while for the opposite direction (15-1) this value is 3189 pax/h. A total observation time of 3 hours has been considered for the application, assuming that hourly passenger flows previously described, preserve the same trend within each hour of the considered period. According to timetable, 30 train

runs have been analyzed for each direction within the observed time interval. Simulation outputs of nominal service (i.e. without any kind of service disturbances) are illustrated in Figure 109 where both train trajectories and on-board passenger flows for each train run and each inter-station track are reported. As can be seen each train run has the same passenger load (in fact in Figure 109b all train loading diagrams are overlapped). Moreover no limits have been set for train capacity (i.e. the max number of passengers that a train can contain). In addition it is immediate to understand that within undisturbed conditions punctuality index referred to train arrivals at station no.15 (for 1-15 direction) and at station no. 1 (for 15-1 direction), is just equal to 100%. Furthermore in this case the total generalized cost estimated over all passengers flows considered during the observed time period is 63893 €.

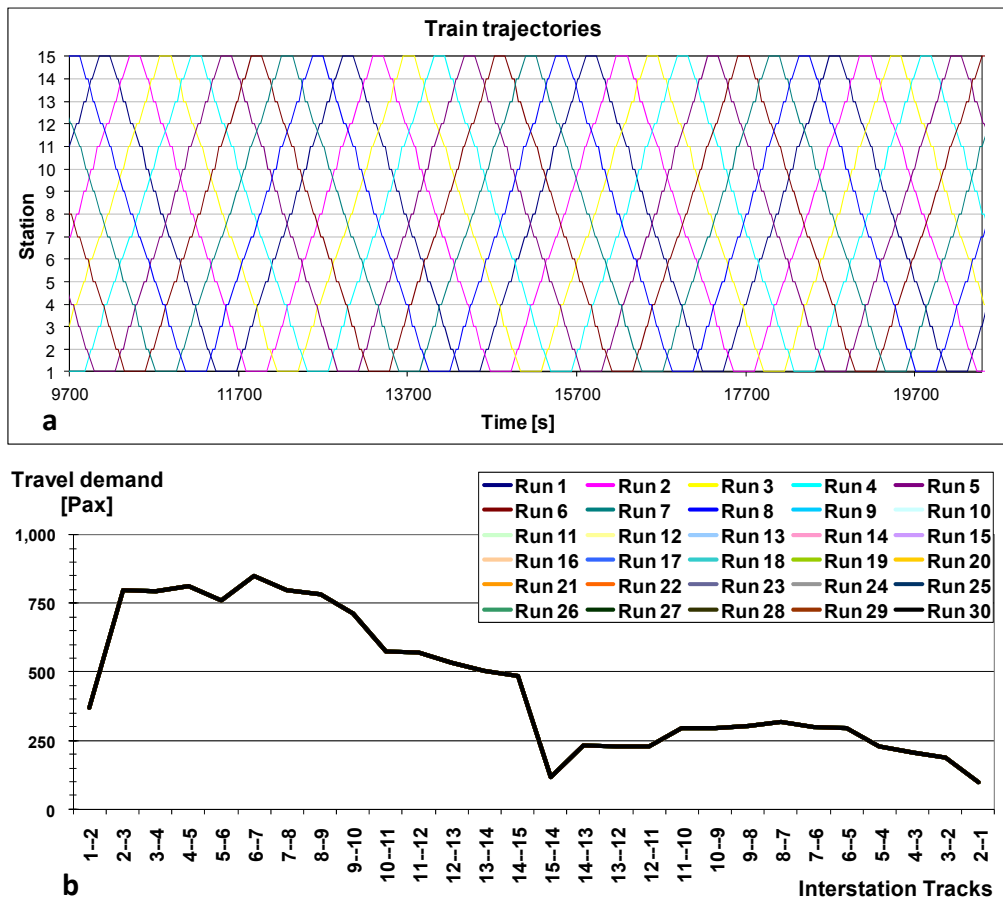


Figure 109. Train trajectories (a) and on-board passenger flows for each train run and each inter-station track (b) within ordinary service.

As said before two different failure scenarios have been considered, supposing that train no. 4 experiences a breakdown while it is performing the second train run along 1-15 direction. In particular such failure causes a reduction of train performances (in terms of

speeds) of 60% in the first failure scenario and 25% in the second one. Then, for both the first and the second failure scenario, the following operation strategies have been analyzed in order to restore normal service:

- *Strategy 1*: one minute after the failure has occurred, a spare from the depot is put on service starting from station no. 1 along direction 1-15, while the corrupted train, although in a degraded state, continues its service until it returns to station no. 1 and enters the depot (where it is stored away).
- *Strategy 2*: the broken train is kept on service until it reaches the pocket track (between stations 8 and 7) where it is stored away. A minute later a spare from the depot is put on service starting from station no. 1 along 1-15 direction.

For each failure scenario the effects induced by each one of the two considered recovery strategies have been assessed calculating both total passengers' generalized cost and the punctuality index at station no. 15 (for 1-15 direction) and no.1 (for 15-1 direction).

Failure scenario 1: Train performance reduction of 60%

In this scenario, the application of operation strategy 1 implies that due to the higher travel time of the corrupted vehicle a strong knock on-delay is suffered from other trains (Figure 110a). For this reason a massive passenger overloading of train runs is observed (Figure 110b). Punctuality index assessed is 65.1%. The total generalized cost estimated for this strategy is 118876 €, and with respect to ordinary service causes an increase of passengers' cost (i.e. a decrease of passengers' satisfaction) of 86%.

The application of recovery strategy 2 instead, implies a strong mitigation of knock-on delays (Figure 111a), since the broken train is removed from the service as soon as possible. But storing away the broken train on the pocket track, all the passengers of the degraded train run (the second run along 1-15 direction) are forced to alight at station 7 and wait for the next run (run 3), which is in fact strongly overloaded (Figure 111b). However in this case the punctuality index is 95.57%, while the total passengers' generalized cost estimated is 65947 € which determines an increase of 3.21% with respect to the normal service.

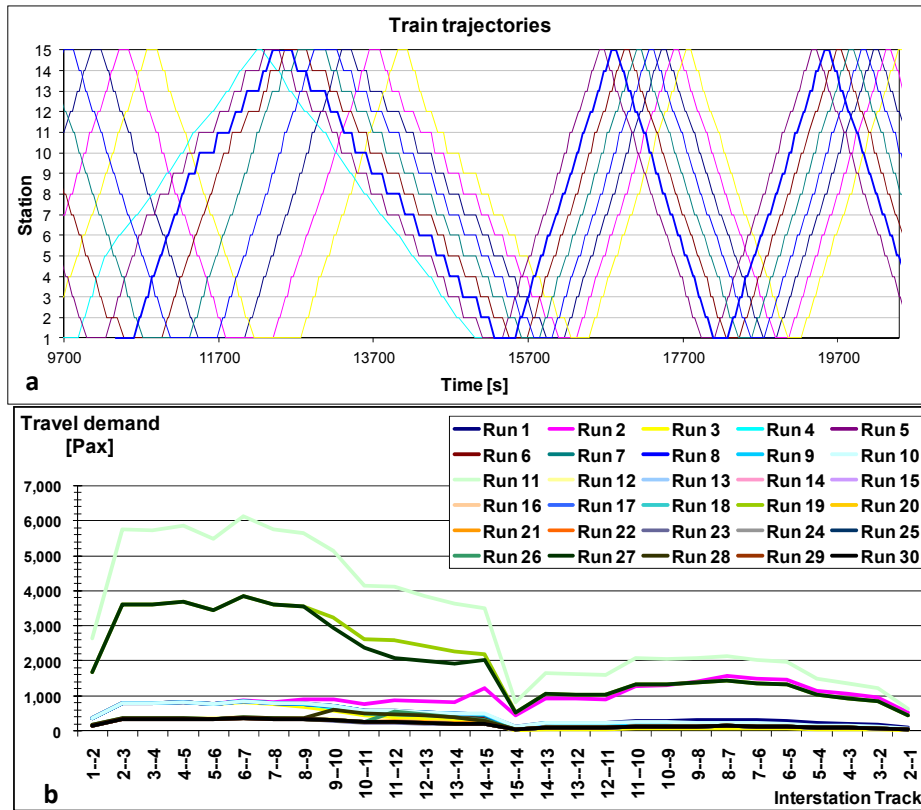


Figure 110. Train trajectories (a) and on-board passenger flows for each train run and each interstation track (b) within Failure scenario 1 and recovery strategy 1.

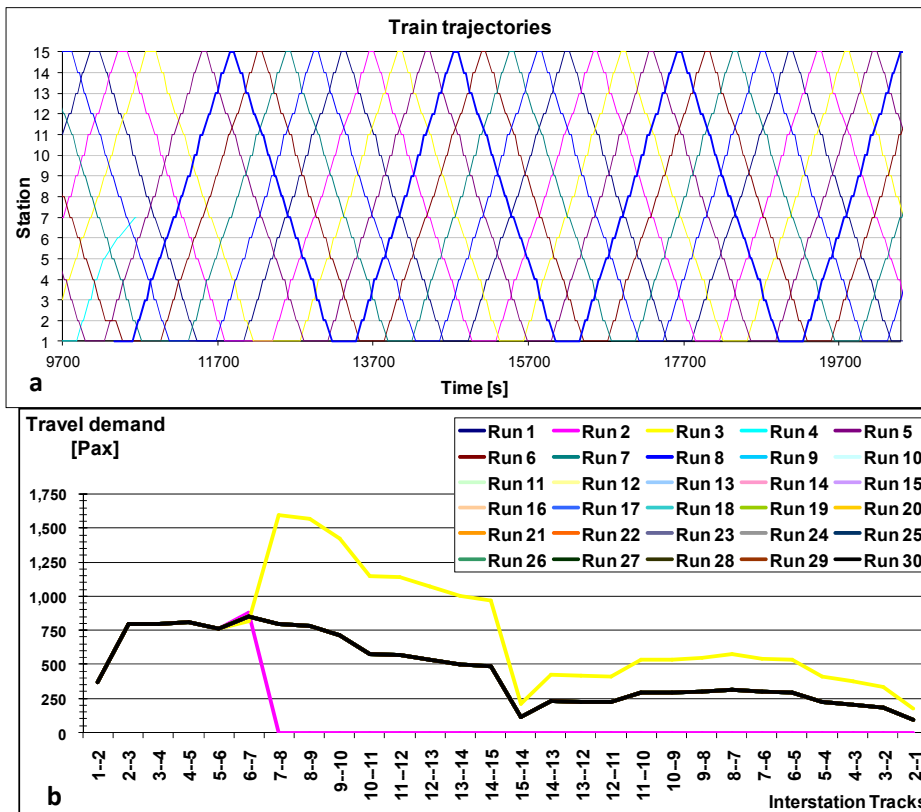


Figure 111. Train trajectories (a) and on-board passenger flows for each train run and each interstation track (b) within Failure scenario 1 and recovery strategy 2.

Failure scenario 2: Train performance reduction of 25%

When train performances of the broken train are only slightly reduced as in this case the application of strategy 1, with respect to the previous scenario causes only light knock on-delays (Figure 112a). In fact as a consequence of this only a soft passenger overloading of train runs is observed (Figure 112b). Moreover the punctuality index assessed for this recovery strategy is 92.64%, while the total generalized cost is 64952 €, causing an increase of passengers' cost (i.e. a decrease of passengers' satisfaction) of only 1.66% with respect to ordinary conditions.

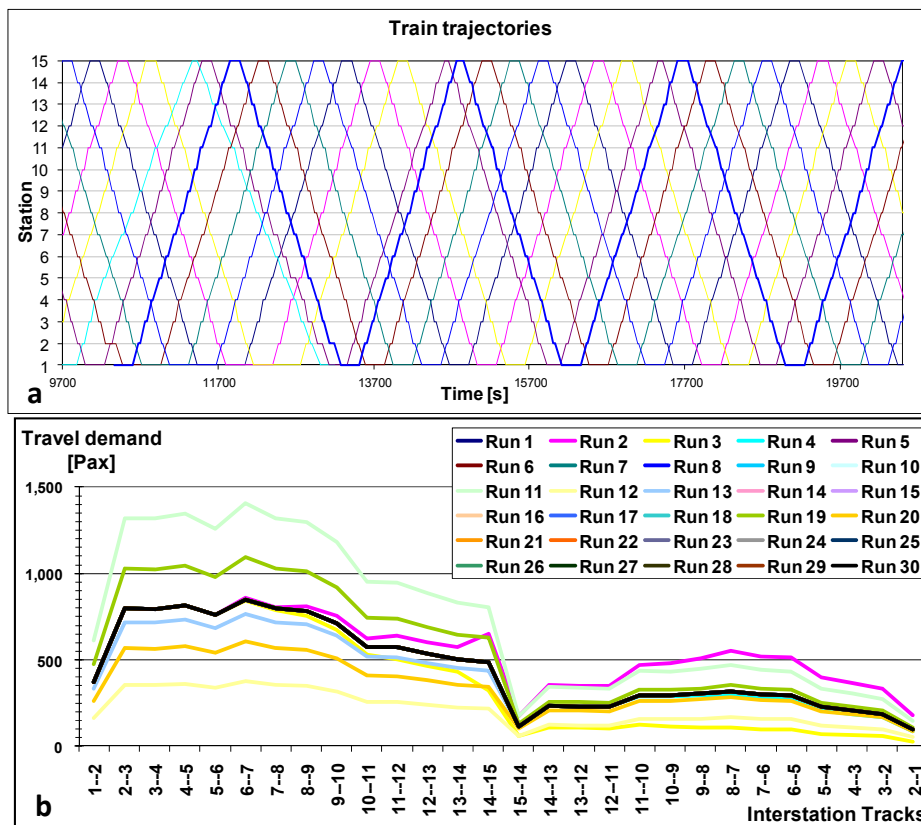


Figure 112. Train trajectories (a) and on-board passenger flows for each train run and each interstation track (b) within Failure scenario 2 and recovery strategy 1.

When applying instead recovery strategy 2 knock-on delays are minimized (Figure 113a), since the broken train is immediately removed from service and the corresponding punctuality index is 95.56%. Although this strategy is the one which gives the best results in terms of punctuality, the total passengers' generalized cost estimated is instead 65605 € which is higher than the one obtained for the previous strategy and determines an increase of 2.68% with respect to normal service (see overloading of Run 3 in Figure 113b).

It is very important to notice that for this failure scenario, strategy 2 induces a lower QoS perceived by passengers, although it is more efficient from the SA (i.e. system performance) point of view. This aspect can be easily explained in practice since when train performances are only slightly reduced as in this case, passengers prefer to remain on the broken train and arrive with at destination with an additional small delay (Strategy 1), rather than get off at station 7 and wait for the next run (Strategy 2). This proves that a design or operational intervention, which presents the highest performance efficiency, not always coincides with the one which assures the highest QoS levels. Therefore the explicit description of both supply and demand-side components as well as their interactions is necessary.

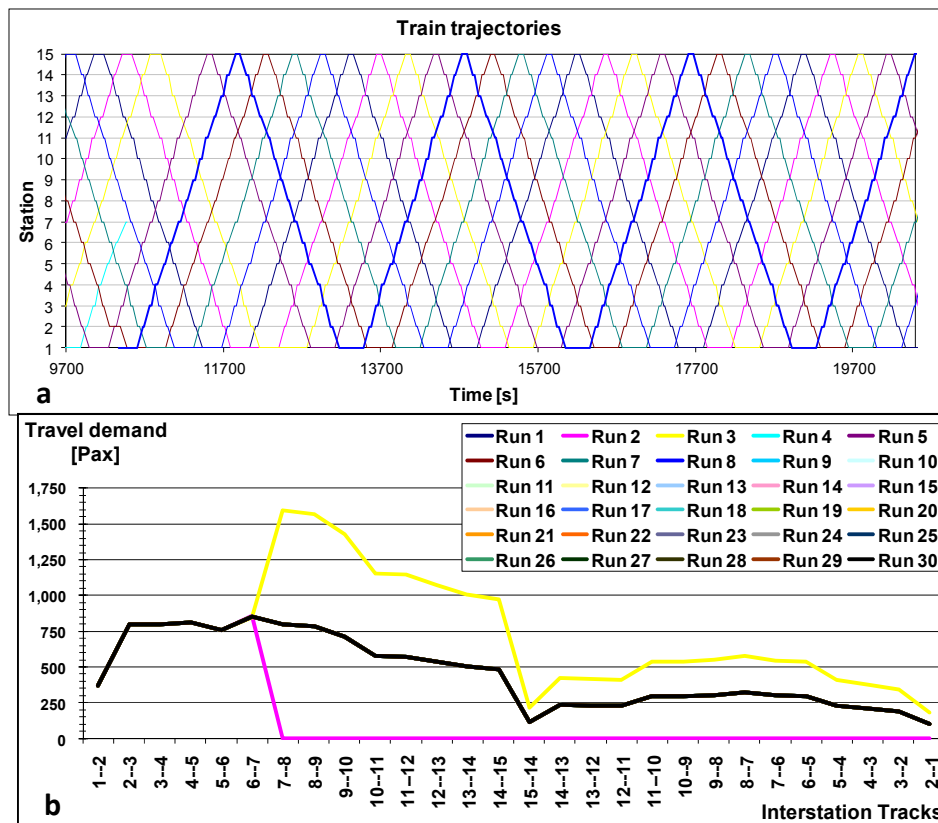


Figure 113. Train trajectories (a) and on-board passenger flows for each train run and each interstation track (b) within Failure scenario 2 and recovery strategy 2.

5.6.2. Towards an integrated simulation framework for Service Availability and Quality of Service evaluation to support RAM analysis

As already seen in the previous sections, railway designers and system suppliers need to perform specific probabilistic analyses (e.g. RAM analysis) to verify if the characteristics of the components supplied (in terms of their reliability and maintainability) as well as the recovery strategies adopted are able to satisfy during the

whole network lifecycle, the requested levels of service availability, as established by customers within contract requirements. Moreover, the introduction of new guidelines at international level (*CER* 2004, *AFNOR* 2006), have highlighted the necessity of measuring and monitoring Quality of Service levels delivered to passengers, in order to assure the achievement of certain standards during operations also when stochastic disruptions to service (e.g. breakdowns, conflicts) do occur. To this purpose, it can be very important for railway designers and suppliers to perform probabilistic analysis which aims at verifying that characteristics of components and recovery strategies considered, not only are able to satisfy SA levels but also QoS standards requested for the whole network lifecycle. To this aim an integrated simulation framework is needed for supporting this kind of activity, which must be able to explicitly simulate both effects on network performances and passenger flows. Furthermore it must be efficient for performing millions Monte-Carlo simulations to draw a significant number of failure events, and accurate to precisely estimate the effects of failures on both SA and QoS. The architecture of such simulation structure has been here identified and its practical (and future) implementation will regard the dynamic cooperation between the microscopic model, the mesoscopic model, as well as the demand assignment module described in this thesis. Since a description of the integration strategy between the microscopic and the mesoscopic model has been already given in the previous section, here a brief illustration of the architecture relative to the proposed integrated framework is carried out (Figure 114).

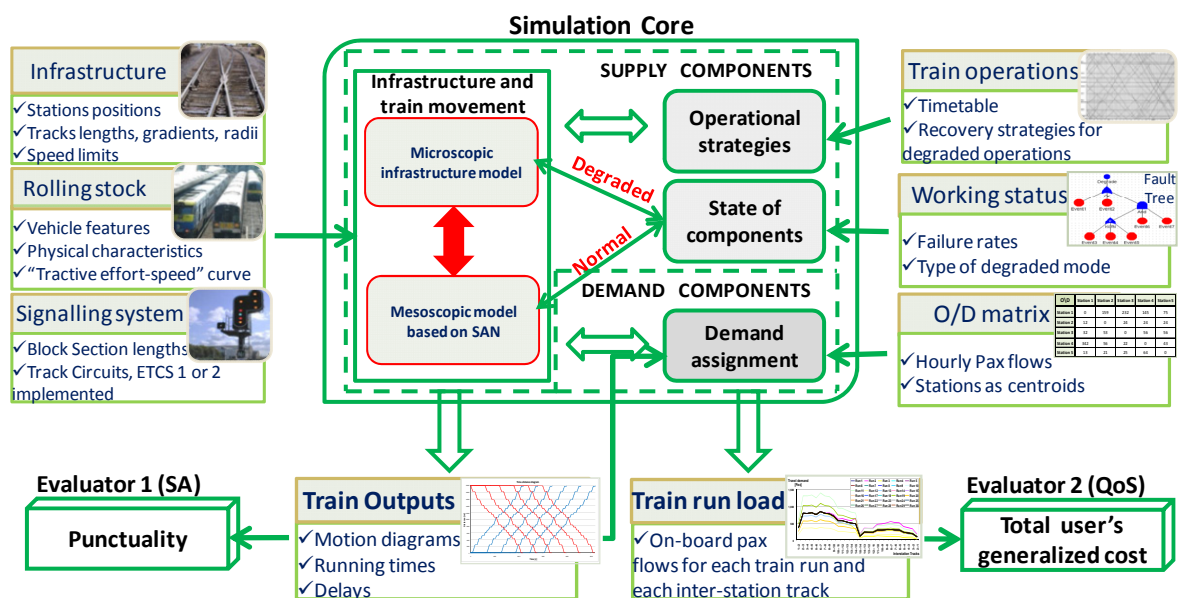


Figure 114. Architecture of the simulation framework for evaluating both SA and QoS.

As can be seen this architecture is composed of two main modules which constitute the simulation core: the “Supply Component” and the “Demand Component” which respectively interact to estimate effects of a certain intervention on both the supply and the demand system. Specifically the “Supply Component” is in turn composed of:

- “*Infrastructure and train movement module*”, which is the main module of the simulation core, and its respective input data regard infrastructure characteristics (positions of stations, length and gradient of rail tracks, track speed limits, type of signalling system, etc.), and physical-mechanical vehicles features (number of coaches and their lengths, max deceleration rate, “tractive effort-speed” curve of the traction unit, etc.). Specifically such module simulates train movements on the track, through interacting with the other modules of the core. As already said, it dynamically integrates the microscopic and the mesoscopic models as illustrated in the previous paragraphs.
- “*Operational strategies module*”. This module considers as input data both operational timetable (scheduled train arrival and departure times at/from each station), and different operation strategies which are addressed to manage ordinary train movements (e.g. train movements towards the depot at the end of the daily service time) as well as degraded conditions to restore normal service after failures of system components. More precisely this module authorizes train movements from stations and block sections, according to timetable and obviously respecting signalling aspects which safely regulate train movements on the track. In case of degraded mode of operation instead, this module activates the simulation of specific operation strategies to recover ordinary conditions after a certain failure event.
- “*Working Status module*”. This module simulates operating states of system components, considering the transition from ordinary to degraded functioning in accordance to failure rates which constitute in fact input data of such module. In particular different degraded operating modes are considered for each component, and state transitions from normal working to one of these degraded states are modelled through the implementation of Markov chains.

In particular the activation of the microscopic model as well as of the mesoscopic model is performed by the “*Working Status*” module. In fact, during ordinary and undisturbed operations, simulation outputs are returned by the mesoscopic model which is the only

one to be activated in this case, in order to exploit its computational efficiency. When instead the “*Working Status*” module draws a stochastic failure event, the microscopic model is activated to accurately evaluate its effects on system performances. Therefore within degraded conditions train outputs are returned by the microscopic model, since only this one is able to provide accurate results in these cases.

Anyway such outputs are used to calculate SA indices (e.g. punctuality, regularity, etc.), and then they are transferred as inputs to the “*Demand assignment*” module for assessing effects on passengers. More precisely this latter module is the only one of the “Demand component” and its features are described below:

- “*Demand assignment module*”. Such module is the responsible for the interaction between passenger demand and railway system. Input data are constituted by train outputs returned by the “*Infrastructure and train movement*” module, as well as origin-destination matrices of rail passenger trips relative to a certain time period and considering station nodes as origin-destination centroids. In particular O-D matrices must be previously split between alighting and boarding passengers flows at stations. Specifically this module determines on-board passengers’ flows for each train run and each inter-station section, through a classical assignment model based on consistency equations.

Outcomes of this module are usefully employed instead to calculate the total user’s generalized cost as a measure of the QoS delivered to passengers.

Therefore as illustrated, the proposed architecture could be an effective tool to support different decisional phases at each level, allowing the explicit simulation of repercussions induced by a certain intervention solution on both the supply and the demand components of railway systems.

Chapter 6. Conclusions

Decisional phases regarding the design of railway infrastructures (station layout, signalling system, etc.) as well as the management of operational strategies (e.g. service timetable, recovery strategies, etc.) need to be supported by simulation models of railway traffic to estimate effects induced by a certain solution on network performances. Typical design activities can regard for example timetable design, analyses for evaluating timetable robustness or network stability, optimal design of signalling layout for maximizing network capacity or minimizing train energy consumption, the identification of effective strategies to manage train movements on marshalling yards, as well as real-time rescheduling of train runs for minimizing effects of disruptions on service. These kind of applications and in general all those activities related to medium-short term designing phases (where infrastructure components and/or operations must be designed in detail) need of very accurate evaluations of impacts on system performances to identify the most effective intervention. To this aim only microscopic infrastructure models can be used to support such phases, since only by representing railway network at a high level of detail, it is possible to precisely describe system behaviour and interactions amongst its components.

Anyway, due to the large amount of input data considered, microscopic models are inefficient for simulating large-sized networks, for performing analyses which require a large amount of model simulations (e.g. probabilistic analyses, “black-box” optimizations, etc.), or for supporting real-time activities (e.g. rescheduling of train runs, management of recovery strategies, etc.). In these cases in fact, practitioners usually prefer to rely on “lower-detailed” (e.g. macroscopic or mesoscopic) or “fixed-speed” models (e.g. models based on alternative graphs: job-shop, etc.) which are more efficient from a computational point of view, but inaccurate in results especially when high congestion levels are on the network.

In addition, the closed-architecture of commercial microscopic models which prevents these models to be customized or interfaced with external applications and mathematical structures (e.g. for probabilistic analysis, or “black-box” optimization), has forced users to develop simulation models built ad hoc for the case-studies considered (therefore devoid of general validity), and responding to certain specific

necessities according to the kind of problems they were dealing with (e.g. *Marinov and Viegas* 2011).

To this aim in this thesis work a “multi-purpose” microscopic infrastructure model of railway systems has been developed to effectively support different design activities, trying to overcome applicability limits of commercial microscopic models.

In particular this work can be split up in two main phases:

1. *A development phase*, in which the microscopic model has been specified, and then practically implemented.
2. *An application phase*, which instead has regarded several applications of the developed model, to solve practical designing problems.

Specifically, the model has been developed in C++ employing an object-oriented technique through which a detailed representation of each system component (e.g. rail vehicles, signalling equipments, rail tracks) has been possible to accurately describe network dynamics as well as interactions amongst its components. Moreover its open-structure allows to be customized, or interfaced with external applications and/or mathematical structures, in order to be “flexible”, in the sense that it can be used for supporting different kinds of design activities. Furthermore it is a synchronous, time-driven model whose architecture is composed of the following four interacting modules, in which input data are administered:

- *Infrastructure module*. Input data required by this module concern with track attributes such as speed limits, gradients, radii, as well as coordinates of infrastructure elements (e.g. positions of stations, junctions). Railway network is here modelled in a link-oriented graph, therefore all track attributes are assigned to links, while coordinates of junctions and stations are assigned to nodes.
- *Rolling stock module*. This module requires as input data all mechanical and physical features of rail vehicles such as the “tractive effort-speed” curve of the traction unit, maximum deceleration rate, jerk value, as well as train composition (number of wagons, their mass, etc.). Train movements are simulated through the integration of the Newton’s motion formula: for each time step, the maximum force between the traction unit’s wheels and the tracks is calculated to determine

acceleration function which is then integrated a first time to provide speed function and a second time to provide train's position. Moreover a module for the calculation of train energy consumption is here included.

- *Signalling system module.* Signalling system is here depicted by an event-driven model. Aspect of signals changes according to the occupation state of the block section that they protect, emulating for a certain kind of signalling technology, the respective working rules performed in the reality. Interactions between vehicles and signalling equipments are modelled, and three different types of signalling systems have been implemented: traditional multi-aspect system, ETCS level 1, ETCS level 2. Hence, signals positions, block section lengths and type of signalling technology are all input of this module.
- *Timetable module.* Departure/arrival times as well as scheduled stop time at stations are input data required by this module. Moreover it is possible to introduce disturbances to ordinary train operations, imposing a deterministic or a stochastic delay to a specific train at a certain station. Furthermore also train dwell times at stations can be modelled as stochastic variables specifying the probability density function, as well as the mean and standard deviation values.

This modular architecture as well as the possibility of initializing each module by specifying relative input data through external files (e.g. text files), let the model be of general validity since it can be applied for whatever case-study.

Furthermore, output data provided by this model are:

- *train diagrams* (e.g. distance-time trajectories, speed-distance diagrams, etc.),
- *train conflicts*, (e.g. conflicts due to block section occupation time overlaps)
- *train statistics* (e.g. arrival/departure delays, punctuality index, etc.)
- *energy consumption diagrams* (e.g. mechanical power-time diagrams, mechanical energy-distance diagrams)

Once the model has been specified and implemented, a parallelization of the simulation process has been carry out, by rewriting part of the code in a concurrent way and employing the parallelization paradigm “OPEN MP” for creating and managing threads.

Tests conducted on a multi-core server, have shown that benefits induced by the parallel architecture on computing times, increase with the number of computer cores and dimensions of the simulation problem, and therefore consent to the model to be efficiently applied also for large-sized networks (i.e. with a diameter > 100 Km) as well as for analyses requiring a large number of model evaluations (e.g. probabilistic analyses, “black-box” optimizations).

A validation phase has been then performed. In particular a first verification of the code has been realized by verifying the congruence of the outputs returned by the model, with those given by a consolidated commercial model *OpenTrack*®, for the same input dataset. Successively a validation process has been carried out by ensuring that simulated train trajectories were consistent with those observed for a real Mass Rapid Transit (MRT) line: the “Cumana” line in Naples city.

Successively a sensitivity analysis has been conducted, to understand how variability (or uncertainty) in model outputs could be apportioned to the variability of model inputs. The Sobol’ “variance-based” technique has been considered, since being a global method consents to investigate homogeneously the whole domain of input variables. Moreover such method allows to assess input sensitivity indices (first-order and total indices) with a number of model simulations that is smaller than the one required by other methods. Results obtained for a real case-study (the “Cumana” line), have consented to identify the design variables which mostly influence a certain network performance. As a consequence, the usefulness of such analysis for supporting design activities has been highlighted, since determining for a certain performance the most relevant parameters, it allows to efficiently allocate economic resources by intervening only on key variables and not also on the other parameters.

As already said, within the second phase of this thesis work, several applications of the model have been realized to support different designing problems. First, a classical “what-if” design approach has been used to evaluate for a real MRT line, effects induced by two different infrastructural interventions on network capacity. Results have confirmed the accuracy of the model in estimating impacts due to a certain solution (e.g. interventions on signalling system, on infrastructure or on operations) and have underlined the usefulness of the model for supporting this kind of decisional phases.

Then the microscopic model has been integrated within a “black-box” optimization loop by interfacing the model itself with the API module of the optimization software “LINDO”. This “black-box” optimization framework has been applied to the real case-study of the “Cumana” line to solve two different problems relative to the design of equi-block signalling layouts. The first problem has regarded the design of an equi-block layout for maximizing the economic efficiency of investment costs. The proposed framework has consented to identify the signalling layout which satisfied the level of capacity required by customers, minimizing investment costs. Outcomes obtained have shown that with respect to the classical criterion used to design signalling layout (which tends to maximize technological efficiency), the proposed design approach allows to strongly reduce investment costs, satisfying the same capacity constraint imposed by customers. Instead the second problem has dealt with the determination of the equi-block signalling layout which allows to reach the best trade-off between user’s satisfaction and investment costs. Such solution in fact would consent to the infrastructure manager to decrease investment costs for installing signalling system without infringing user’s satisfaction. Results have illustrated how this design solution can lead to consistent reductions of signalling costs by only increasing the required service headway of a value which is unperceived by passengers. In addition the effectiveness of the “black-box” optimization framework developed, has been underlined for solving this kind of decisional problems.

A further application has regarded the field of the so-called RAM (*Reliability, Availability, Maintainability*) analysis, in which millions Monte-Carlo simulations are needed to verify if the required service availability indices can be satisfied by using components with certain values of reliability and maintainability. In particular a comparison with an efficient “event-driven” mesoscopic model (which considers component failure rates as inputs) based on the Stochastic Activity Network formalism has been conducted to understand the trade-offs between efficiency of this latter and accuracy of the microscopic model developed. Differences of the two models in terms of computing efficiency and result accuracy have been quantified through their application to a simple metro line. Results underlined applicability limits of the two models (inaccuracy of the mesoscopic and inefficiency of the microscopic) especially for congested network, highlighted the necessity of dynamically integrate the approaches to effectively support RAM analysis.

A modelling strategy to integrate the two models has been identified and the corresponding architecture has been defined. Specifically an initial microscopic simulation of ordinary service is launched to calculate train free-flow running times. Then these running times are transferred as inputs to the mesoscopic model, which in turn is activated to perform millions Monte Carlo simulations of ordinary service, in order to draw stochastic failure events (according to components failure rates). Only when a failure event is drawn, the microscopic model is activated to accurately estimate its effect on network performances.

Furthermore the microscopic model has been interfaced with a module for simulating passenger demand, in order to evaluate impacts of a certain intervention not only on Service Availability (SA) but also on Quality of Service (QoS) offered to passengers, as requested by recent international guidelines (introduced by European bodies like the *CER*). This simulation framework has been applied to a metro case-study, where two different failure scenarios have been analyzed and for each one of them, two recovery strategies have been compared. Results illustrate that not always the most efficient solution in terms of network performances (SA) guarantees the highest QoS perceived by passengers, and highlight the necessity of explicitly simulate the effects on both railway operations and passenger travel demand. This aspect, hence underlines the relevance of the introduced simulation framework to support decisional phases at each level.

Moreover the architecture of a more complex simulation framework has been defined to take into account also failure rates of system components and therefore impacts of stochastic failure events on both SA and QoS. To this purpose the dynamic cooperation of the microscopic model, the “event-driven” mesoscopic model and the demand assignment module will be implemented. In fact this dynamic integration, will allow to perform effective RAM analyses for making inference on effects of stochastic failures not only on network performances but also on demand. In particular, the simulation of undisturbed ordinary conditions will be realized by the mesoscopic model, whose computing efficiency consents to perform millions Monte Carlo simulations in order to draw stochastic failure events. Then only when a failure event is drawn, the microscopic model is activated to accurately estimate train performances during degraded operation. Train outputs (returned by the mesoscopic model during ordinary service and by the microscopic one during degraded conditions) are successively employed to calculate SA

indices (e.g. punctuality, regularity, etc.). After that such outputs are transferred as inputs to the demand assignment module for assessing effects on passengers.

Future research will be addressed to determine an effective modelling strategy which allows the dynamic integration between the microscopic and the mesoscopic approach, assuring not only the dynamic exchange of input/output variables between the models, but also the automatic activation of both the models when their intervention is required.

Furthermore both the microscopic and the mesoscopic models will be dynamically interfaced with the module for the simulation of passenger flows, in order to guarantee an automatic transfer of input/output data between the infrastructure-side and the demand-side simulation models. Such dynamic cooperation will consent in fact the employment of such integrated architecture also for performing probabilistic analyses and/or for supporting real-time decisional phases.

References

Albrecht T., Energy-efficient train operation, in: Hansen I.A., Pachl J. (Editors), *Railway Timetable and Traffic: Analysis-Modelling-Simulation*, Eurail Press, 2008.

Albrecht, T., Energy-efficient train control in suburban railways: experiences gained from onboard tests of a driver assistance system, *Proceedings of the 1st International seminar of Railway and Operations Modelling and Analysis*, Delft, the Netherlands, 2005.

Association of Train Operating Companies, *Passenger Demand Forecasting Handbook*, London, 2005.

Batley R., Dargay J., Wardman M., The Impacts Of Lateness And Reliability On Passenger Rail Demand, *Transportation Research Part E*, Vol. 47, pp. 61-72, 2011.

Brunger O., Dahlhaus E., Running Time Estimation, in: Hansen I.A., Pachl J. (Editors), *Railway Timetable and Traffic: Analysis-Modelling-Simulation*, Eurail Press, 2008.

Carey M., Carville S., Testing Schedule Performance and Reliability for Train stations, *Journal of the Operational Research Society*, Vol. 51, pp. 666-682.

Carey M., Kwiecinski A., Stochastic Approximation to the Effects of Headways on Knock-on Delays of Trains, *Transportation Research Part B*, Vol. 28, pp. 251-267, 1994.

Cascetta E., *Transportation System Analyses, Models And Applications*, Springer, New York, 2009.

CENELEC, Railway applications – Specification and demonstration of reliability, availability, maintainability and safety (RAMS), *EN 50126*, 1999.

CER, Implementation of the Charter on Rail Passenger Services in Europe, Progress Report, November, 2004.

Chang C.S., Du D., Further Improvement Of Optimization Method For Mass Transit Signaling Block-Layout Design Using Differential Evolution, *IEE Proceedings, Electrical Power Applications*, Vol. 146, No. 5, pp.559-569, September 1999.

Chang C.S., Du D., Improved Optimization Method Using Genetic Algorithms For Mass Transit Signaling Block-Layout Design, *IEE Proceedings, Electrical Power Applications*, Vol. 145, No. 3, pp. 266-272, May 1998.

Ciuffo B., Punzo V., Quaglietta E., Kriging Meta-Modelling to Verify Traffic Micro-Simulation Calibration Methods, Proceeding of the Transportation Research Board Annual Meeting, Washington DC, 23-27 January 2011.

References

- Corman, F., D'Ariano A., Hansen I. A., Evaluating Disturbante Robustness Of Railway Dispatching Measures, *Proceedings of the 2nd International Conference on Models and Technologies for ITS*, Leuven, Belgium, 22-24 June 2011.
- Cukier, R.I., Fortuin, C.M., Schuler, K.E., Petscheck, A.G., Schailbly, J.H., Study Of The Sensitivity Of Coupled Reaction Systems To Uncertainties In Rate Coefficients. *The Journal of Chemical Physics*, Vol. 26, pp. 1-42, 1973.
- Curtius E., Kniffler A., Neue Erkenntnisse über die Haftung Zwischen Treibrad Und Schiene, *Elektrische Bahnen*, Vol. 9 pp. 201-210, 1943.
- D'Ariano A., Albrecht T., Running Time Re-Optimization During Real-Time Timetable Perturbations, *Computers in Railways X*, pp.531-540, 2006.
- D'Ariano A., Pranzo M., An Advanced Real-Time Train Dispatching System For Minimizing The Propagation Of Delays In A Dispatching Area Under Severe Disturbances, *Proceedings of the 2nd International Seminar On Railway Operations Modelling and Analysis*, Hannover, 2007.
- D'Ariano A., Pranzo M., Hansen I. A., Conflict Resolution and Train Speed Coordination for solving Real-Time Timetable Perturbations, *IEEE Transactions on Intelligent Transportation Systems*, Vol.8, NO. 2, June 2007.
- Di Febraro A., Giua A., Sistemi ad Eventi Discreti, McGraw-Hill, 2002, Milan, Italy.
- Dijkstra E.W., A Note On Two Problems In Connexion With Graphs, *Numerische Mathematik*, Vol. 1, pp. 269-271, 1959.
- European Rail Research Advisory Council, *Strategic Rail Research Agenda 2020*, 2007, Available: <http://www.errac.org/IMG/pdf/SRRA-2007.pdf> (last access 15.03.2011).
- Garstenauer K., Appel B., Marktentwicklung für ERTMS-Lösungen in Europa und Übersee, *ETR-Eisenbahntechnische Rundschau*, Vol. 11, pp. 666-668, 2007.
- Gill D.C., Goodman C. J., Computer-Based Optimization Techniques For Mass Transit Railway Signaling Design, *IEE Proceedings B, Electrical Power Applications*, Vol.139, No. 3, pp. 261-275, 1992.
- Goverde R., Punctuality of Railway Operations and Timetable Stability Analysis, TRAIL Thesis Series No. T2005/10, Delft, 2005
- Goverde R., Railway Timetable Stability Analysis Using Max-Plus System Theory, *Transportation Research Part B*, Vol. 41, pp.179-201, 2007.
- Gröger, T., Simulation Der Fahrplanerstellung Auf Der Basis Eines Hierarchischen Trassenmanagements Und Nachweis Der Stabilität Der Betriebsabwicklung Dissertation, *Veröffentlichungen des Verkehrswissenschaftlichen Institutes der RWTH Aachen*, Vol. 60, 2002.

Guilloux, J. P., Das Signalsystem der Hochgeschwindigkeitsstrecken in Frankreich, in: *Signal + Draht*, Vol. 2, EurailPress, 1990.

Hansen I.A., Improving Railway Punctuality by Automatic Piloting, *Proceedings of the IEEE Intelligent transportation Systems Conference*, pp. 792-797, Oakland, USA, 2001.

Hansen I.A., Pachl J., Railway Timetable and Traffic, Eurail Press, Hamburg, Germany, 2008.

Hansen, I. Increase of Capacity Through Optimised Timetabling, *Computers in Railway IX*, WIT Press, pp. 529-538, 2004.

Happel, O., Sperrzeiten als Grundlage für die Fahrplankonstruktion, *Eisenbahntechnische Rundschau*, Vol. 8, No. 2, pp. 79-90, 1959.

Hauptmann, D., Automatiche und Diskriminierungsfreie Ermittlung Von Fahrplantrassen in Beliebigen Großen Netzen Spurgeführter Verkehrssysteme (Automatic and Non-Discriminatory Train Path Allocation In Railway Networks Of Arbitrary Size), Dissertation am Institut für Verkehrswesen, Eisenbahnbau und betrieb (Schriftenreihe No. 54), Universität Hannover, Hestra Verlag, 2000.

Ho T.K., Mao B.H., Yuan Z.Z., Liu H.D., Fung Y.F., Computer simulation and Modeling in Railway Applications, *Computer Physics Communications*, Vol. 143, pp. 1-10, 2002.

Homma, T., Saltelli, A., Importance Measures In Global Sensitivity Analysis Of Nonlinear Models. *Reliability Engineering and System Safety*, Vol. 52, pp. 1-17, 1996.

Howlett P.G., Pudney P.J., Energy Efficient Train Control, *Advances in Industrial Control*, Springer, London 1995.

Huisman T., Boucherie R., Running Time on Railway Sections with Heterogeneous Train Traffic, *Transportation Research Part B*, Vol. 35, pp. 271-292, 2001.

Jacobs J. Rechnergestützte Konfliktermittlung und Entscheidungsunterstützung bei der Disposition des Zuglaufs, Ph.D. Thesis, Veröffentlichungen des Verkehrswissenschaftlichen Institutes der RWTH Aachen, Heft 61, 2003.

Jacobs J., Reducing Delays by Means of Computer-Aided “On-The-Spot” Rescheduling, *Computer In Railways IX*, pp. 603-612, 2004.

Jacobs J., Rescheduling, in: Hansen I.A., Pachl J. (Editors), *Railway Timetable and Traffic: Analysis-Modelling-Simulation*, Eurail Press, 2008.

Jacques, J. Lavergne, C., Devictor, N., Analysis In Presence Of Model Uncertainties And Correlated Inputs. *Reliability Engineering and System Safety*, vol. 91, pp. 1126-1134, 2006.

- Jansen M.J.W., Analysis Of Variance Designs For Model Output, *Computer Physics Communications*, Vol. 117, pp. 35–43, 1999.
- Jansen M.J.W., Rossing W.A.H., Daamen R.A., Monte Carlo Estimation Of Uncertainty Contributions From Several Independent Multivariate Sources, in: J. Gasmanand, G. van Straten (Eds.), *Predictability and Nonlinear Modelling in Natural Sciences and Economics*, Kluwer Academic Publishers, Dordrecht, , pp. 334–343, 1994.
- Ke B. R., Chen M. C., Lin C. L., Block-Layout Design Using Max-Min Ant System For Saving Energy On Mass Rapid Transit Systems, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 10, No. 2, pp. 226-235, June 2009.
- Kettner M., Netz-Evaluation und Engpassbehandlung mit Makrosjopischen Modellen des Eisenbahnbetriebs (Network Evaluation And Bottleneck Handling with Macroscopic Models for Railway Operation), Dissertation am Institut für Verkehrswesen, *Eisenbahnbau und Betrieb* (Schriftenreihe No. 65), Universität Hannover, EurailPress, 2005.
- Kettner M., Prinz R., Sewcyk B., NEMO – Netz – Evaluations-Modell Bei Den OBB, *Eisenbahntechnische Rundschau (ETR)*, Vol. 3, pp.117-121, 2001.
- Kettner M., Sewcyk B., A Model For Transportation Planning And Railway Network Evaluation, *Proceedings of the 9th World Congress on Intelligent Transport Systems*, Chicago, USA, October 14-17, 2002.
- Kettner M., Sewcyk B., Eickmann C., Integrating Microscopic And Macroscopic Models For Railway Network Evaluation, *Proceedings of the European Transport Conference*, Strasbourg, France, 8-10 October 2003.
- Khmelnitsky E., On an Optimal Control Problem of Train Operation, *IEEE Transaction on Automatic Control*, Vol. 45, No. 7, March 2000.
- Kondo R., Introduction of Advanced Type Automatic Train Stop System, in: *Japanese Railway Engineering*, Vol. 4, 1980.
- Kotler Ph., Marketing Management, Analysis, Planning, Implementation and Control, Prentice Hall, 1991.
- Koutsopoulos H., Wang Z., Simulation of Urban Rail Operations: Application Framework, *Transportation Research Record: Journal of the Transportation Research Board*, No. 2006, pp 84-91, 2007.
- Kroon L., Dekker R., Vromans M., Cyclic Railway Timetabling: a Stochastic Optimization Approach, in: Geraets F., Kroon L., Schobel A., Wagner D., Zaroliagis C., (editors): *Algorithmic Methods for Railway Optimization*, Lecture Notes in Computer Science, Springer, Berlin, 2007.

- Kroon L., Peeters L., A Variable Trip Time Model for Cyclic Railway Timetabling, *Transportation Science*, Vol. 37, pp. 198-212, 2003.
- Landex A., Evaluation of Railway Networks with Single Track Operation Using the UIC 406 Capacity Method, *Networks and Spatial Economics*, Vol. 9, pp. 7-23, 2009.
- Law A.M., Kelton W.D., Simulation modeling and analysis, McGraw-Hill, 2000.
- Law, A.M., Simulation Modelling and Analysis. Fourth Edition. McGraw-Hill, New York, 2007.
- Luethi M., Weidmann U.A., Laube F. B., Medeossi G., Rescheduling and Train Control: A New Framework for Railroad Traffic Control in Heavily Used Networks, *Proceedings of the Transportation Research Board 86th Annual Meeting*, Washington DC, January 21-25, 2007.
- Marinov M., Viegas J., A Mesoscopic Simulation Modelling Methodology For Analyzing And Evaluating Freight Train Operations In A Rail Network, *Simulation Modelling Practice and Theory*, Vol. 19, pp.516-539, 2011.
- Martinez, B., Vitoriano, A., Fernandez, A., Cucala, A.P., Statistical Dwell Time Model For Metro Lines, *WIT Transactions on the Built Environment*, Vol. 96, pp. 223-232, 2007.
- Mascis A., Pacciarelli D., Job Shop Scheduling With Blocking And No-Wait Constraints, *European Journal of Operational Research*, vol. 143, no. 3, pagg. 498-517, Dec. 2002.
- Mazzarello M., Ottaviani E., A Traffic Management System For Real-Time Traffic Optimization In Railways, *Transportation Research Part B*, 41, pagg. 246-274, 2007.
- Mazzeo A., Mazzocca N., Nardone R., Quaglietta E., D'Acerno L., Punzo V., Montella B., Lamberti I., Marmo P., An Integrated Approach For Availability And Qos Evaluation In Railway Systems, *Proceedings of the 30th International Conference on Computer Safety, Reliability and Security (SafeComp)*, Naples, 19-22 September 2011.
- Medeossi, G., Huerlimann D., Longo G., Stochastic Micro-Simulation As A Timetable Robustness Estimation Tool, *Proceedings of the 3rd International Seminar on Railway Research (IAROR)*, Zurich, Switzerland, 2009.
- Middelkoop, D., Bouwmann, M., Testing the Stability of The Rail Network, *Computers in Railways VII*, pp. 995-1002, WIT Press, Southampton, 2002.
- Middelkoop, D., Bouwmann, M., Train Network Simulator for Support of Network Wide Planning of Infrastructure and Timetables, *Computers in Railways VII*, pp. 267-276. WIT Press, Southampton, 2000.
- Muller C. Th., Kinematik, Spurführungsgeometrie und Führungsvermögen der Eisenbahntrasse, *Glaser's Annalen*, Vol. 77, pp. 264-281, 1953.

References

- Muller, C. Th., Wear Profiles Of Wheels And Rails, Office of Research and Experiment (ORE) of the International Union of Railways (UIC), ORE-Report C9/RP6, Utrecht, 1960.
- Murali P., Dessouky M.M., Ordonez F., Palmer K., A Delay Estimation Technique for Single and Double-Track Railroads, *Transportation Research Part E*, Vol. 46, pp. 483-495, 2010.
- Nash A., Huerlimann D., Railroad Simulation Using Open-Track, *Computers in Railways IX*, WIT Press, Southampton, UK, 2004.
- Nathanail E., Measuring The Quality Of Service For Passenger On The Hellenic Railways, *Transportation Research Part A*, Vol. 42, pp. 48-66, 2008.
- Nordeen, M., Stability Analysis of Cyclic Timetables for a Highly Interconnected Rail Network, PhD Thesis, Swiss Federal Institute of Technology (EPFL), Lausanne, 1996.
- Pachl J., Modelling Specific Signalling Features in Computer-Based Scheduling Systems, *Rail Delft Proceedings of the International Association of Railway Operation Research*, Delft, 2005.
- Pachl J., Railway Operation And Control, VTD Rail Publishing, Mountlake Terrace, 2002.
- Perticaroli F., Sistemi Elettrici Per i Trasporti: Trazione Elettrica, Casa Editrice Ambrosiana, 2001
- Qi, Z., Baoming, H., Dewei, L., Modelling and Simulation of Passenger Alighting and Boarding Movement in Beijing Metro Stations, *Transportation Research Part C*, Vol. 16, pp. 635-649, 2008.
- Quaglietta E., D'Acierno L., Punzo V., Nardone R., Mazzocca N., A Simulation Framework For Supporting Design And Real-Time Decisional Phases In Railway Systems, *Proceedings of the 14th IEEE ITS Conference*, Washington DC, 5-7 October 2011.
- Quaglietta E., Punzo V., A Global Sensitivity Analysis Framework For Supporting The Design Of Railway Systems, *Proceedings of the 91st TRB Annual Meeting*, Washington DC, 22-26 January 2012.
- Quaglietta E., Punzo V., Montella B., Nardone R., Mazzocca N., Towards A Hybrid Mesoscopic-Microscopic Railway Simulation Model, *Proceedings of the 2nd international Conference on Models and Technologies for ITS*, Lowen, Belgium, 22-24 June 2011.
- Radtke A., EDV-Verfahren zur Modellierung des Eisenbahnbetriebs (Software Tools to Model the Railway Operation), Habilitation am Institut für Verkehrswesen,

Eisenbahnbau und Betrieb (Schriftenreihe No. 64), Universitat Hannover, EurailPress, 2005.

Radtke A., Infrastructure modelling, in: Hansen I.A., Pachl J. (Editors), *Railway Timetable and Traffic: Analysis-Modelling-Simulation*, Eurail Press, 2008.

Radtke A., Watson R., Railway simulation in the United Kingdom, RTR No.1, pp.24-28, 2007.

Railway Technical Web Pages, Available at: <http://www.railway-technical.com> (last access 16.11.2011).

Retiveau R., La Signalisation Ferroviaire, Presses de l'Ecole Nationale des Ponts et Chaussees, Paris, 1987.

Sakowitz C., Wendler E., Optimising Train Priorities to Support The Regulation of Train Service With the Assistance of Active or Deductive Databases, *Computers in Railway X*, pp. 489-499, 2006.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., Tarantola, S., Variance Based Sensitivity Analysis Of Model Output. Design And Estimator For The Total Sensitivity Index. *Computer Physics Communications*, Vol. 181, pp. 259-270, 2010.

Saltelli, A., Making Best Use Of Model Evaluations To Compute Sensitivity Indices. *Computer Physics Communications*, Vol. 145, pp. 280-297, 2002.

Saltelli, A., Ratto, M., Anres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S., Global Sensitivity Analysis , The Primer - John Wiley and Sons Ltd, Cheichester England. 2008

Saltelli, A., Tarantola, S., Campolongo, F., Ratto, M., Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models. John Wiley and Sons Ltd, Cheichester England. 2004

Saltelli, A., Tarantola, S., On The Relative Importance Of Input Factors In Mathematical Models: Safety Assessment For Nuclear Waste Disposal. *Journal of American Statistical Association*, vol. 97, pp. 702- 709, 2002.

Sauthoff F., Die Bewegungswiderstande der Eisenbahnwagen unter Besonderer Beruck-Sichtigung der Neueren Versuche der Deutschen Reichsbahn, Ph.D. Thesis, Technical University, Berlin, 1932.

Schumacher A., Ein Hybrides Verfahren zur Umlaufplanung von Fahrzeugen des Spurgefuhrten Verkehrs (A Hybrid Method for Vehicle Allocation in Rail Traffic), Dissertation am Institut fur Verkehrswesen, *Eisenbahnbau und Betrieb* (Schriftenreihe No. 60), Universitat Hannover, EurailPress, 2004.

Serafini P., Ukovich W., A Mathematical Model for Periodic Event Scheduling Problems, *SIAM Journal of Discrete Mathematics*, Vol. 2, pp. 550-581, 1989.

References

Sewcyk B., Makroskopische Abbildung des Eisenbahnbetriebs in Modellen Zur Langfristigen Infrastrukturplanung (Macroscopic display of Railway Operation in Models for Long-Term Infrastructure Planning). Dissertation am Institut für Verkehrswesen, *Eisenbahnbau und Betrieb* (Schriftenreihe No. 61), Universität Hannover, EurailPress, 2004.

Sewcyk B., Radtke A., Wilfinger G., Combining Microscopic and Macroscopic Infrastructure Planning Models, *Proceedings IAROR*, Hannover, 2007.

Shultze, K., Modell Für Die Asynchrone Simulation Des Betriebes In Teilen Des Eisenbahnnetzes Dissertation, *Veröffentlichungen des Verkehrswissenschaftlichen Institutes der RWTH Aachen*, Vol. 38, 1985.

Siefer T., Radtke A., Railway Simulation, Key for Operation and Optimal Use, *Proceedings of the 1st International seminar of Railway and Operations Modelling and Analysis*, Delft, the Netherlands, June 8-10, 2005.

Sobol, I.M., Sensitivity Analysis For Non-Linear Mathematical Models. *Mathematical models and Computation Experiments*, Vol.1, pp. 407-414, 1993.

Sobol, I.M., Uniformly Distributed Sequences With An Additional Uniform Property. *USSR Computational Mathematics and Mathematical Physics*, Vol. 16 (5), pp. 236-242, 1976.

Sobol', I.M., Tarantola, S., Gatelli, D., Kucherenko, S., Mauntz, W., Estimating The Approximation Error When Fixing Unessential Factors In Global Sensitivity Analysis. *Reliability Engineering and System Safety*, Vol. 92, pp. 957-960, 2007.

Strahl, Die Berechnung der Fahrzeiten und Geschwindigkeiten von Eisenbahnzügen aus den Belastungsgrenzen der Lokomotiven, *Glaser's Annalen für Gewerbe und Bauwesen*, Vol. 73, pp. 869-871, 1913.

Such W.H., Mechanical and Electrical Interlocking, IRSE Green Booklet No. 3, 2nd Edition, 1956.

Theeg G., Vlasenko S., Railway Signalling and Interlocking International Compendium, Eurail Press, 2009

UIC, Code 406 – Capacity. 1st Edition, 2004

Watanabe I., Ushijima Y., Fukuda M., Takashige T., Development of Digital ATC System, in: *Quarterly Report of RTRI*, Vol. 1, 1999.

Wegele S., Schnieder E., Dispatching of Train Operations Using Genetic Algorithms, *Proceedings of the 9th International Conference on Computer-Aided Scheduling of Public Transport*-CD-ROM, San Diego, 9-11 August, 2004.

Wendler E., the Scheduled Time On Railway Lines, *Transportation Research Part B*, Vol. 41, pp. 148-158, 2007.

References

- Wendler, E., ETCS und Kapazität, *Proceedings of the VDE Kongress*, Vol. 2, pp. 369-374, Aachen, 2006.
- White T., Elements of Train Dispatching, VTD Rail Publishing, Mountlake Terrace, 2003.
- Winter P., Compendium on ERTMS, DVV Media/EurailPress, Hamburg, 2009.
- Yalcinkaya O., Bayhan G.M., Modelling and Optimization of Average Travel Time for A Metro Line by Simulation And Response And Response Surface Methodolgy, *European Journal of Operational Research*, Vol. 196, pp. 225-233, 2009.
- Yamanouchi S., Safety and ATC of Shinkansen, *Japanese Railway Engineering*, Vol. 2, 1979.
- Yuan J., Goverde R., Hansen I.A., Propagation of Train Delays in Stations, in Allan J., et al. (Editors), *Computers In Railways VIII*, WIT Press, pp. 975-984, Southampton, 2002.
- Yuan J., Hansen I.A., Optimizing Capacity Utilization of Stations by Estimating Knock-On Delays, *Transportation Research Part B*, Vol. 41, pp. 202-217, 2007.
- Yuan, J., Stochastic Modelling of Train Delays and Delay Propagation in Stations, PhD Thesis, Delft University of Technology, 2006.